INTERFACE DESIGN FOR THE SUPERVISORY CONTROL OF MULTIPLE HETEROGENEOUS UNMANNED VEHICLES

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

By

KYLE JOSEPH BEHYMER
M.S., Wright State University, 2005
B.A., Thomas More College, 2002

2017
Wright State University
I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Kyle Joseph Behymer ENTITLED Interface Design for the Supervisory Control of Multiple Heterogeneous Unmanned Vehicles BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

John M Flach, Ph.D.
Dissertation Director

Scott N. J. Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Final Examination

Gary N. Burns, Ph.D.

Ion Juvina, Ph.D.

Thomas C. Hughes, Ph.D.

Robert E. W. Fyffe, Ph.D.
Vice President for Research and Dean, Graduate School
The Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT) described within this dissertation were developed by a triservice effort through the Assistant Secretary of Defense for Research and Engineering (ASD/R&E) Autonomy Research Pilot Initiative “Realizing Autonomy via Intelligent Adaptive Hybrid Control”. IMPACT development was led by the Air Force Research Laboratory’s 711th Human Performance Wing Supervisory Control and Cognition Branch (711HPW/RHCI). An in-depth description of the entire IMPACT project will be published as an AFRL technical report (AFRL-RH-WP-TR-2017-TBD).

The views expressed in this dissertation are those of the author and do not necessarily reflect the official policy or position of the Air Force, the Department of Defense, or the United States Government.

ABSTRACT

Behymer, Kyle Joseph. Ph.D., Department of Psychology, Wright State University, 2017. Interface Design for the Supervisory Control of Multiple Heterogeneous Unmanned Vehicles.

In order to meet the demand for enabling one operator to control multiple heterogeneous unmanned vehicles numerous automated support systems are being developed. These systems are too often focused on replacing, rather than supporting, the human decision maker. In contrast, the Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT) system was designed from a collaborative systems approach that allowed human operators to work with autonomous systems to accomplish mission tasks. Multiple cognitive task analyses were conducted with base defense experts as well as unmanned vehicles (UV) operators to inform the development of human-autonomy interfaces (HAI) that were designed to support an operator’s skill-based, rule-based, and knowledge-based behaviors using ecological interface design principles. This research describes the development and empirical evaluation of the IMPACT HAI using a synthetic task environment in which participants used twelve UVs to support base defense operations. A 2 X 2 within-participants experimental design was used to compare IMPACT’s HAI to a Baseline HAI condition across two levels of mission complexity. Eight participants completed four hour-long trials in which they were responsible for responding to incoming mission tasks.
Participants both preferred and performed better with the IMPACT HAI as compared to the Baseline HAI. These results suggest that ecological interface design principles can be used to generate user interface concepts that not only support skill-based behaviors, but also rule-based and knowledge-based behaviors.
# TABLE OF CONTENTS

CHAPTER I. Introduction .................................................................................................. 1
   The Domain: Unmanned Vehicle Supervisory Control .................................................. 1
   From Autonomous Systems to Sociotechnical Systems: Designing Effective
   Collaborations ................................................................................................................. 3
   The Prosthetic/Substitution Approach: The Technical Limits .......................................... 8
   The Prosthetic/Substitution Approach: Unintended Social Consequences ...................... 11
   A Collaborative Systems Approach: Complementing Capabilities ................................. 14
   A Collaborative Systems Approach: Creating Social Cohesion ..................................... 19
   Summary ....................................................................................................................... 22

CHAPTER II. Understand ................................................................................................. 25
   Cognitive Task Analysis ............................................................................................... 26
   The Domain: Base Defense ........................................................................................... 28
   Cognitive Task Analysis: Search and Rescue ............................................................... 29
   Cognitive Task Analysis: Base Defense ....................................................................... 32
   Cognitive Task Analysis: UAV and UGV Operations ................................................... 34
   Cognitive Task Analysis: Informing Autonomous Capabilities ..................................... 35
   Summary ....................................................................................................................... 37

CHAPTER III. Generate ................................................................................................... 39
   The Art of Design .......................................................................................................... 40
   Practice .......................................................................................................................... 43
   Preparing for Serendipity .............................................................................................. 48
   Spanning the Gulf of Execution. Rule-Based Interaction. Direct Manipulation .............. 51
   Spanning the Gulf of Evaluation. Direct Perception ..................................................... 62
   Supporting Knowledge-Based Behavior .......................................................................... 71
   IMPACT Vignette ......................................................................................................... 76
   Summary ....................................................................................................................... 83

CHAPTER IV. Validate ..................................................................................................... 84
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Tyranny of the Quantifiable</td>
<td>85</td>
</tr>
<tr>
<td>Context Matters</td>
<td>89</td>
</tr>
<tr>
<td>Evaluating IMPACT’s HAI</td>
<td>92</td>
</tr>
<tr>
<td>CHAPTER V. Experimental Study Methods</td>
<td>95</td>
</tr>
<tr>
<td>Participants</td>
<td>95</td>
</tr>
<tr>
<td>Design</td>
<td>95</td>
</tr>
<tr>
<td>Apparatus</td>
<td>98</td>
</tr>
<tr>
<td>Scenario</td>
<td>101</td>
</tr>
<tr>
<td>Procedure</td>
<td>104</td>
</tr>
<tr>
<td>Dependent Measures</td>
<td>108</td>
</tr>
<tr>
<td>CHAPTER VI. Results</td>
<td>110</td>
</tr>
<tr>
<td>IMPACT vs. Baseline Subjective Measures</td>
<td>110</td>
</tr>
<tr>
<td>IMPACT vs. Baseline Objective Measures</td>
<td>115</td>
</tr>
<tr>
<td>IMPACT vs. Baseline Summary</td>
<td>130</td>
</tr>
<tr>
<td>Qualitative Results</td>
<td>132</td>
</tr>
<tr>
<td>CHAPTER VII. Discussion</td>
<td>137</td>
</tr>
<tr>
<td>Limitations</td>
<td>137</td>
</tr>
<tr>
<td>Scientific Contribution</td>
<td>139</td>
</tr>
<tr>
<td>Future Research</td>
<td>140</td>
</tr>
<tr>
<td>Conclusion</td>
<td>141</td>
</tr>
<tr>
<td>AFTERWORD</td>
<td>146</td>
</tr>
<tr>
<td>CHAPTER VIII. References</td>
<td>148</td>
</tr>
<tr>
<td>CHAPTER IX. Appendices</td>
<td>159</td>
</tr>
<tr>
<td>APPENDIX A. Acronym List</td>
<td>160</td>
</tr>
<tr>
<td>APPENDIX B. Random Anti-Terror Measures Lists</td>
<td>162</td>
</tr>
<tr>
<td>APPENDIX C. Sample Sensor Operator Script</td>
<td>163</td>
</tr>
<tr>
<td>APPENDIX D. Sample TOC Operator Script</td>
<td>164</td>
</tr>
<tr>
<td>APPENDIX E. Background Questionnaire</td>
<td>165</td>
</tr>
<tr>
<td>APPENDIX F. Base Map</td>
<td>166</td>
</tr>
</tbody>
</table>
APPENDIX G. UV Comparison Chart................................................................. 167
APPENDIX H. Sample Scoresheet................................................................. 168
APPENDIX I. Overall System Questionnaire.................................................. 170
APPENDIX J. System Usability Scale (Brooke, 1996)......................................... 171
APPENDIX K. IMPACT vs. Baseline Questionnaire ......................................... 173
LIST OF FIGURES

Figure 1. Framework for Human-Machine teams collaborating to solve a complex domain problem. ................................................................................................................. 6
Figure 2. The Gulfs of Execution and Evaluation (adapted from Hutchins, Hollan, and Norman, 1985). ................................................................................................................. 15
Figure 3. Optimal Search Pattern Decision Tree. ......................................................................................................................... 30
Figure 4. Decision Process for Suspicious Vehicle. ......................................................................................................................... 36
Figure 5. Decision Process for Suspicious Vehicle with Autonomy. ......................................................................................................................... 37
Figure 6. A Notional Playbook® interface (Shaw et al., 2010). ......................................................................................................................... 55
Figure 7. City of Heroes Mastermind Interface (NCSOFT, 2012. Author’s Screenshot). 56
Figure 8. Mastermind Interface. (NCSOFT, 2012. Author’s Screenshot). ......................................................................................................................... 59
Figure 9. IMPACT Play Calling Interface ......................................................................................................................... 60
Figure 10. Right Click Play Menu. ......................................................................................................................................................... 61
Figure 11. Task Manager ......................................................................................................................................................... 62
Figure 12. Process Algebra (Varricchio, Chaudhari, & Frazzoli, 2014). ......................................................................................................................... 63
Figure 13. Visual Thinking Sprocket (Findler, 2011). ......................................................................................................................... 64
Figure 14. Representing Play Quality (Behymer, Ruff, Mersch, & Calhoun, 2015). ......................................................................................................................... 65
Figure 15. Play Quality Bar Chart (Behymer, Ruff, Mersch, & Calhoun, 2015). ......................................................................................................................... 66
Figure 16. A Pareto Front ......................................................................................................................................................... 67
Figure 17. Comparing USVs across Many Parameters (Savitz, et al., 2013). ......................................................................................................................... 68
Figure 18. The COA Planner. ......................................................................................................................................................... 69
Figure 19. Active Play Manager ......................................................................................................................................................... 71
Figure 20. Medium Level of Cardiac Risk (McEwen, 2012). ......................................................................................................................................................... 72
Figure 21. Low Level of Cardiac Risk (McEwen, 2012). ......................................................................................................................................................... 73
Figure 22. Play Workbook ......................................................................................................................................................... 74
Figure 23. Chat Message Alert of an Unidentified Watercraft. ......................................................................................................................... 76
Figure 24. Task Manager Provides QRC for Unidentified Watercraft. ......................................................................................................................... 77
Figure 25. Plan for Responding to Unidentified Watercraft, FN-43 to Point Sierra. ......................................................................................................................... 78
Figure 26. FN-43 Flies to Point Sierra to Investigate the Unidentified Watercraft. ......................................................................................................................... 79
Figure 27. FN-43’s Sensor Fails. ......................................................................................................................................................... 80
Figure 28. COA Planner Compares Plans to get to Point Sierra. ......................................................................................................................... 80
Figure 29. Vignette Flowchart. ......................................................................................................................................................... 82
Figure 30. Wason Task (1966). ......................................................................................................................................................... 87
Figure 31. Modified Wason Task (Cosmides & Tooby, 1992). ......................................................................................................................................................... 88
Figure 32. Modified Wason Task (Cosmides & Tooby, 1992). ......................................................................................................................................................... 88
Figure 33. Factors Leading to the Shootdown. (Snook, 2000). ......................................................................................................................................................... 90
Figure 34. Abstraction Space (Flach and Voorhorst, 2016). ......................................................................................................................................................... 91
Figure 35. Abstraction Space (Snook, 2000). ......................................................................................................................................................... 92
Figure 36. UxAS Wizard. ................................................................. 97
Figure 37. Experimental Configuration. ................................................. 99
Figure 38. C2 Station Layout for both IMPACT and Baseline Conditions. 101
Figure 39. Comparison of Overall Ratings for IMPACT and Baseline .......... 111
Figure 40. Comparison of System Usability Scale (Brooke, 1996) ............... 112
Figure 41. Comparison of SUS Score for IMPACT and Baseline. ............... 112
Figure 42. IMPACT vs. Baseline Ratings for Mission Tasks ...................... 113
Figure 43. NASA-TLX Results. ...................................................... 115
Figure 44. Percentage of RAMs Completed Correctly ............................ 116
Figure 45. Normal Base Defense Events Outcome Measure ..................... 118
Figure 46. Normal Base Defense Events Process Measure ....................... 119
Figure 47. Intruder Events Outcome Measure. ...................................... 120
Figure 48. Intruder Event Process Measure. .......................................... 121
Figure 49. Percentage of System Failures and Environmental Events Responded to Correctly. .................................................. 123
Figure 50. Commander Query Accuracy. ........................................... 124
Figure 51. Commander Query Response Time. ..................................... 125
Figure 52. Commander Query Response Time vs. Process Score. .............. 127
Figure 53. Mean Time from Play Initiation to Play Execution. .................... 128
Figure 54. Mean Number of Clicks from Play Initiation to Play Execution. ... 129
Figure 55. Picking a Closer, But Slower UV. ....................................... 133
Figure 56. Time to Arrive. .................................................................. 134
LIST OF TABLES

Table 1. Chess Elo Ratings. ................................................................................................ 4
Table 2. Recipe Differences highlighted in gray. ............................................................. 18
Table 3. Commander’s Intent. (Shattuck 2000). .............................................................. 57
Table 4. IMPACT and Baseline Differences. ................................................................. 96
Table 5. Task Complexity. ........................................................................................... 98
Table 6. IMPACT vs. Baseline Participant Comments. ............................................... 114
Table 7. IMPACT vs. Baseline Summary. ................................................................. 131
ACKNOWLEDGEMENTS

There’s a popular proverb, made famous by Hillary Clinton, that it takes a village to raise a child. Well, after 15 years of working on this I can say that it also takes a village to complete a dissertation. I’d like to thank some of the villagers who have helped me along the way.

I’m continuously amazed at the dedication, creativity, empathy, and effort of those who enter the education profession, and I thank all the educators who have taught me along the way. My first grade teacher Andrea Lonneman helped set me on the path to becoming a voracious reader and my high school English teacher, Sharon Meyerrose taught me how to write.

I had a wonderful quartet of professors in my undergraduate psychology department at Thomas More College. Dr. Kathie Langen and Dr. Lawrence Boehm helped me explore all the various fields within psychology and helped me discover human factors. Dr. William Porter taught me how much fun psychology could be and ensured I would always have at least one intelligent thing to say about every famous psychologist, from A to Z (Adler to Zimbardo). My undergraduate advisor Dr. Maria McLean taught me what it means to be a research scientist, how to write scientific papers, and how to communicate effectively. She is an exemplar of what a psychology professor should be. I wish I could have brought home a Psych Bowl win for them. Thomas Hext taught me how to program, and made sure that my brain wasn’t just fertilizer for my hair.

My dissertation advisor, Dr. John Flach, is the smartest person I’ve ever met. I learn something new every time we speak and I’m amazed at how he is constantly learning. He’s been a tireless advocate for me over the years, and has never let me forget about finishing my dissertation no matter how many distractions life presented. Without his encouragement and persistence, I would have given up on completing my dissertation years ago. That being said, the Southeastern Conference is better at football than the Big Ten and I hope John realizes that someday.

In addition to being a member of my dissertation committee Dr. Thomas Hughes is also my boss. Tom’s unwavering support during this process was critical and his insights vastly improved this work. I was lucky enough to meet my additional committee members Dr. Ion Juvina and Dr. Gary Burns during this project, and both gave freely of their time and expertise to improve this dissertation.

Special thanks to the 711th Human Performance Wing Supervisory Control and Cognition Branch (711HPW/RHCI) for allowing me to use IMPACT as part of my dissertation. Dr.
Mark Draper, IMPACT’s program manager, is the visionary behind IMPACT and he graciously allowed me to play in the sandbox he and his team created.

The interface concepts described within were developed by IMPACT’s Human-Machine interface team, led by Gloria Calhoun. Gloria went above and beyond for me during this project, and I’ve yet to meet anyone who works harder at their job than she does at hers. I’d also like to thank other HAI team members especially Elizabeth Frost, Patrick Dudenhofer, and Capt Chad Breeden.

The cognitive task analyses, experimental design, and experiment were conducted under the auspices of Dr. Michael Patzek, IMPACT’s test and evaluation team lead. Mike provided expert guidance and support throughout the research effort. Clayton Rothwell played a key role in the early stages of the project, assisting in CTAs, designing scenarios, and running participants. Clayton also owes me $20. Dakota Evans played the sensor operator when needed and helped design the experiment. Jessie Bartik was a late addition to the project, but essential, tracking down potential participants and stepping in as the sensor operator.

Special thanks to Heath Ruff, who as a member of both IMPACT’s HAI team and test and evaluation team had to deal with me 40 hours a week, every week. Heath stepped in to help test IMPACT when no one else wanted to and spent countless hours in the lab. Heath manned the test operator console during the study and was of invaluable help during the data analysis process. Most importantly, Heath Ruff is the most cantankerous person I’ve ever met and I hope he never changes. I owe him several proper breakfasts.

The Fusion software team (Allen Rowe, Dr. Daylond Hooper, Sarah Spriggs, George Bearden, Mike Howard, Adam Buchanan, and Bruce Clay to name but a few) worked tirelessly and constantly for years to make sure IMPACT worked flawlessly. Kenneth Wickline was instrumental in making sure the data captured was the data needed and managed all the configurations.

Dr. Derek Kingston, lead developer of IMPACT’s cooperative control algorithm, was very gracious with his time and was always willing to make software changes to improve the experiment. The intelligent agent team led by Dr. Scott Douglass were a pleasure to work with, and IMPACT’s success is due in no small part to their willingness to work closely with the HAI team. Special thanks to Dr. Mike Hansen for bringing the freestyle chess tournament to my attention.

Dr. Mei-Hua Lin was my graduate school buddy and a constant source of encouragement. Watching Dr. April Courtice defend her dissertation showed me a path forward when hope was lost. Brian McKenna would be my first round draft pick whether it’s for
human factors psychology or intramural sports and is always willing to take a look at whatever crazy idea is currently occupying my mind.

Thanks to my Dad, who has always been my hero. Thanks to my Mom, who is more responsible than anyone for making me who I am today. I hope I’ve made them proud. Thanks to my little sister Lesleigh, who I’m very proud of. Thanks to my children, James, Audrey, and Evelyn, for understanding when Daddy needed to work on his dissertation. Finally, thanks to my wife Andrea, my first reader and my last, for everything, always.
DEDICATION

To Andrea: Each time you happen to me all over again.
CHAPTER I. Introduction

The Domain: Unmanned Vehicle Supervisory Control

Unmanned vehicles (UVs) are being developed and integrated into both the military and commercial world at an ever-increasing rate. The United States Department of Defense’s (DoD) unmanned aerial vehicle (UAV) inventory grew more than 40-fold from 2002 to 2010 (Gertler, 2012) and Amazon has announced its goal of using UAVs to deliver packages (CBS News, 2013). This growth in the use of UVs is not limited to the air, with unmanned ground vehicles (UGVs) being developed for use on land, and unmanned surface vehicles (USVs) and unmanned underwater vehicles (UUVs) being developed for use above and below the sea (Quincy, Thompson, Moran, Nilsson, & Johnson, 2004).

One of the key challenges facing UV developers is reducing the manpower burden associated with operating UVs (United States Department of Defense, 2013). A typical Predator UAV, for example, is operated by three individuals, a pilot, a Sensor Operator (SO), and a Mission Intelligence Coordinator (MIC) (United States Air Force, 2010). The pilot is responsible for flying the aircraft and placing the aircraft in the optimal position for the SO. The SO is responsible for operating the sensor to collect intelligence data for the unit they are supporting. The MIC is responsible for coordinating between the supported unit, the Predator crew, and the Distributed Common
Ground System (DCGS), which is an even larger team of individuals tasked with exploiting the data the Predator collects (United States Air Force, 2009).

The DoD wants to flip this control paradigm; rather than multiple operators controlling one UV, the goal is to have one operator control multiple UVs with the assistance of automated systems (United States Department of Defense, 2013). In this paradigm an automated system essentially flies, drives, or steers the UVs (Stanard et al., 2011) while the human operator focuses on assigning high-level tasks, monitoring the situation, and adjusting to unexpected events (Spriggs, Warfield, Calhoun, & Ruff, 2010). A significant challenge facing automated system designers is designing a user interface that not only allows the human operator to intervene in flying, driving, or steering the UV when necessary, but also provides the operator with tools to address task level concerns such as maintaining surveillance on a specific location (Uhrmann & Schulte, 2011).

The Office of the Assistant Secretary of Defense for Research and Engineering is currently funding a tri-service (Air Force, Army, and Navy) research program to develop Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT). IMPACT’s goal is to integrate autonomous technologies like cooperative control algorithms (CCAs) that plan optimal routes, intelligent agents (IAs) that recommend optimal courses of action, and autonomic frameworks that monitor ongoing events with innovative human-autonomy interface (HAI) concepts to assist human operators in controlling multiple UVs via high level commands called plays. (Behymer,
Rothwell, Ruff, Patzek, Calhoun, Draper, Douglass, Kingston, & Lange, 2017). Much like a head coach calling a play that his or her team then executes, the UV operator can call a play that his or her UVs then execute.

Autonomous technologies like CCAs, IAs, and autonomics frameworks are critical in enabling a single operator to control multiple UVs. However, these technologies are often developed in isolation, with research focused on improving the capability of the technology without considering how the technology will interact with the human operator. The following section examines the limitations of this approach and suggests an alternative approach to designing HAIss that support effective human-autonomy collaborations.

**From Autonomous Systems to Sociotechnical Systems: Designing Effective Collaborations**

In 2005, Playchess.com hosted a chess tournament in which teams of human players could use computer assistance during matches\(^1\). The chess super computer Hydra was also entered into the competition, and after recently defeating Grand Master Michael Adams 5 ½ - ½ in a six game match, was considered to be the prohibitive favorite. Surprisingly, Hydra was eliminated before the semi-finals, with three of the four semi-finalists consisting of Grand Master led teams equipped with supercomputers. Even more

---

\(^1\) Various names have been proposed for this type of chess, including advanced chess, cyborg chess, centaur chess, and freestyle chess.
surprising was the fourth semi-finalist and eventual winner, team ZachS, composed of two relatively amateur chess players named Steven Crampton and Zackary Stephen using ordinary computers (Thompson, 2014).

The Elo Rating system\(^2\) — a method of rating chess player skill level based on head to head results — puts team ZachS’s victory into perspective. Table 1 lists Elo ratings ranging from novice to world champion. Current world champion Magnus Carlsen obtained the highest Elo rating (2882) in history for a human player (Garry Kasparov’s best was 2851, Bobby Fischer’s was 2785) (Chessgames.com Statistics Page, n.d.). At the time of the tournament, Hydra’s estimated Elo rating was 3000, and the runner up team was led by two 2600+ Grand Masters. Crampton and Stephen’s Elo ratings were 1685 and 1398 respectively (Dark horse ZackS wins Freestyle Chess Tournament, 2005).

Table 1. Chess Elo Ratings.

<table>
<thead>
<tr>
<th>Elo</th>
<th>Skill Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1200</td>
<td>Novice</td>
</tr>
<tr>
<td>2000</td>
<td>Expert</td>
</tr>
<tr>
<td>2400</td>
<td>Master</td>
</tr>
<tr>
<td>2600</td>
<td>Grand Master</td>
</tr>
<tr>
<td>2700</td>
<td>World Champion</td>
</tr>
</tbody>
</table>

Team ZachS was vastly outclassed in chess skill and computer hardware, yet overcame Hydra and the Grand Masters armed with super computers by quickly and

\(^2\) The Elo Rating system, developed for chess by Arpad Elo (1978), has also been used to measure skill in many sports including baseball, basketball, football, soccer, and tennis.
efficiently manipulating their machines to deeply explore relevant positions and shrink
the search space for their chess computers (Kasparov, 2010). The higher skill level of
Hydra and the Grand Masters equipped with super computers was not enough to
overcome the seamless collaboration between the less skilled amateurs and their weaker
computers. As Garry Kasparov stated, “Weak human + machine + better process was
superior to a strong computer alone and, more remarkably, superior to a strong human +
machine + inferior process” (Kasparov, 2010).

In fact, the human-machine combination has the potential to outperform human-
alone and computer-alone in many domains. For example, human forecasters at the
National Weather Service can improve the accuracy of computer precipitation forecasts
by 25% and computer temperature forecasts by 10% over computer-only forecasts
(Silver, 2012), and human-computer teams have the potential to outperform both doctors
and computer algorithms at correctly interpreting mammograms (Gaynor, Wyner, &
Gupta, 2014).

However, as the chess example illustrates, group performance is more than the
sum of the abilities of the individuals that compose the group. For example, the collective
intelligence of a group of people is more highly correlated with the group’s social
sensitivity, equality in turn-taking, and the number of women in the group than the
average intelligence of group members or the IQ of the group’s smartest person (Wooley
et al., 2010). Similarly, cardiac surgery efficiency is more dependent on the surgical
team’s cumulative experience than the individual experience of the attending surgeon (Elbardissi et al., 2013). To express this concept as an old sports adage, *teamwork beats talent when talent doesn’t work as a team*. Pairing the best human with the best computer won’t necessarily result in the best performance.

The human-machine team is like a pair of scissors cutting through the fabric of the work domain. Sharpening either of the blades — increasing the capabilities of either the human or the machine — might lead to cleaner cuts. But if there were no hinge — no effective interface — to hold the blades together, no matter how sharp the blades were, the scissors would not cut at all.

Figure 1. Framework for Human-Machine teams collaborating to solve a complex domain problem.

As shown in Figure 1, the distributed sociotechnical system includes multiple potential loops, where each loop is limited in terms of access to information or perception (P) (i.e., what it knows) and action or control capability (C) (i.e., what it can do). The
quality of the observer and control processes depend on the quality of coupling across the multiple loops. If the coupling is rich, then the sociotechnical system can be a better observer than any of the components. For example, each loop may provide unique information about the state of the problem domain and redundant data across the loops can be useful in filtering sampling noise (averaging, common mode rejection). Also, rich coupling allows coordination of multiple actions to achieve a common goal. Without coupling, the actions of each loop will be a potential disturbance relative to other loops. If the coupling within the network of collaborating agents is rich — if there is effective communication — then the whole can serve as a more effective control system than any of the components. If the coupling within the network is poor, then there is a potential for the whole to be worse than the best component due to interference between loops.

Unfortunately, system developers often focus on increasing the capabilities of autonomous agents, without giving sufficient consideration to how they will interface with human operators. This approach often fails to recognize the technical limitations of autonomous components and the potential of a human-machine team. Additionally, focusing on improving the technical capabilities of autonomous agents without considering how these will interact with human operators often leads to poor coupling within the human-machine team. This dissertation proposes an alternative approach — frame the problem as interfacing to the problem domain and the other ‘agents’, not only to improve observability and controllability, but also to take the technical and social aspects involved in enriching the coupling between components into account.
The Prosthetic/Substitution Approach: The Technical Limits

Replacing the human user with autonomous systems, or at the very least, designing autonomous systems to compensate for or overcome the limitations of the human user has been referred to as a prosthesis approach (Roth, Bennett, & Woods, 1987) or substitution-based approach (Hollnagel, 1999), and is based on the idea that designers should identify human weaknesses and replace them with automation strengths (Dekker & Woods, 2002). The origin of this approach can be traced to a 1955 Dartmouth manifesto in which a group of artificial intelligence (AI) scientists — John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon — proposed a goal of discovering how machines could solve the kinds of problems that had previously been the domain of skilled humans, without considering if/how these machines would interact with people. (Epstein, 2015). The question wasn’t how to design an autonomous system that could collaborate with a person to complete a task; rather, the question was how to design an autonomous system that could substitute for human capabilities (Klein et al., 2004).

Advocates of the prosthetic/substitution approach often present humans as poor decision makers, citing studies in which human participants in contrived laboratory tasks perform poorly compared to mathematical decision-making models like Bayes’ Theorem (Flach & Voorhorst, 2016). These studies conclude that human rationality is bounded
and is therefore limited. What is often underrepresented is the fact that autonomous systems are bounded as well.

In 2005, the Defense Advanced Research Project Agency (DARPA) created a research program called COORDINATORS, whose goal was to develop hand-held automated agents that would help geographically distributed warfighters coordinate and adapt mission plans in response to unexpected events (Kohout, 2011). In 2008, a capstone exercise\(^3\) was conducted to compare two different automated agent approaches — developed by two separate teams — with the performance of a control team of human operators (Maheswaran et al., 2009). One automated agent approach removed the human from the loop entirely; the team of humans it supported acted as actuators, only taking actions the agent assigned to them. This approach performed significantly worse than the control team of human operators, and the developers concluded that if automated agents are not provided with appropriate situation constraints, they will inevitably trend towards a subpar solution in the face of a highly dynamic environment (Barbuiescu et al., 2010). The second automated agent approach fared slightly better, but only because its design allowed its developer team to input a human-devised strategy tailored to the specific scenario prior to the exercise. According to the designers of this approach, the automated agents still failed because they did not have an effective method of narrowing the

\(^3\) While employed at JXT Applications, INC., the author managed a DARPA Phase II STTR (Small Business Technology Transfer) that developed the user interface used by one of the automated agent developer teams during this capstone exercise.
enormous search space of the exercise’s dynamic environment. The designers’ key takeaway is worth quoting verbatim: “The most interesting result of the evaluation is that it is so difficult to outscore humans in a complex planning and scheduling problem.” (Maheswaran et al., 2009).

When an autonomous system is presented as the answer to the “problem” of human bounded rationality it is inevitable that the technology will eventually reach its own limits. Consider Watson, the IBM supercomputer designed to compete in Jeopardy! — a television quiz show in which three contestants compete to earn the most money by answering trivia questions. After two rounds, Watson was soundly defeating two of the best human Jeopardy! players ever, Ken Jennings and Brad Rutter⁴, by a score of $36,681 to $2,400 and $5,400 respectively. Then came Final Jeopardy! The category was U.S. Cities, and the clue was “Its largest airport is named for a World War II hero; its second largest, for a World War II battle”. Jennings and Ritter correctly answered “What is Chicago?”, while Watson answered “What is Toronto??????”⁵.

Watson’s response elicited an audible groan from an audience full of IBM programmers likely thinking, “Toronto isn’t a U.S. city.” Except, as Watson was all too aware, Toronto is a U.S. city — in Illinois, Indiana, Iowa, Kansas, Missouri, Ohio, and South Dakota. Adding to the confusion, Watson was programmed to deemphasize

---

⁴ Jennings won a record 74 straight games in 2004. Rutter has never been defeated by a human player in Jeopardy! and is not only the all-time Jeopardy! money winner but also the all-time game show money winner.

⁵ The number of question marks indicates the lack of confidence Watson had in its answer.
category names, as they are often only weakly tied to the content of the clue, and can contain puns or other forms of wordplay. So while this question was relatively easy for a human player, it proved to be Watson’s Achilles’ heel.

In the haste to replace irrational humans with rational machines, advocates of the prosthetic/substitution approach have failed to recognize that autonomous systems also have limits — what’s more, they are overlooking a better solution. Imagine a Warfighter teaming with a DARPA COORDINATOR agent. Imagine Brad Rutter teaming with Watson to play *Jeopardy!*. The critical point is that the rationality of all agents — human and machine — are bounded with respect to the complexity of many work domains. Thus, it will often be necessary to combine the capabilities of multiple agents, each with unique bounds and capabilities, in order to meet the demands for effective performance reflected in Ashby’s Law of Requisite Variety (Ashby, 1956).

*The Prosthetic/Substitution Approach: Unintended Social Consequences*


Thirty-five years ago Weiner and Curry (1980) noted that the general public had two opinions in regard to automation: skepticism about its capabilities and fear of its consequences — widespread unemployment at best, and Orwellian dystopia at worst. The prosthetic/substitution approach has done little to alter these opinions in subsequent
years, with coverage in the media being divided between fear mongering and disdain (Epstein, 2015). The story of John Henry’s epic battle and ultimately Pyrrhic victory over the steam engine exemplifies the fear that people have of being replaced (or even destroyed) by automation, a fear that is omnipresent in popular culture. The first cinema robot appeared in Fritz Lang’s 1927 silent film *Metropolis*, a *Maschinenmensch* (German for machine-human) created by the evil scientist Rotwang to replace Maria, an activist working to better the lives of the workers on whose backs the gleaming city of Metropolis has been built. The Maschinenmensch is designed to look exactly like Maria and has a single goal: to destroy Maria’s reputation among the workers. The Maschinenmensch sows chaos among the workers and they riot, causing floods and destroying parts of the city. Eventually the subterfuge is discovered and the Maschinenmensch is burned at the stake. Similar themes are present in modern films such as *Terminator*, *The Matrix*, and *Avengers: Age of Ultron* — a machine designed to replace humanity turns on humanity. Both Schafer et al., (2015) and Parasuraman and Riley (1997) have argued that these fictional portrayals have influenced society’s perception of autonomous systems and may create dissonance when people interact with autonomous systems in the real world.

Another unintended social consequence of failing to take social factors into account when designing automated systems is disdain. In 1993, Microsoft started the Lumiere project, with the goal of developing an automated capability that could detect a user’s goals based on their actions and provide assistance to the user to meet his or her
goals (Horvitz, n.d.). In 1997, after more than 25,000 hours were spent on usability testing, Clippy\textsuperscript{6} was released as part of Office 97. It was so despised that its removal from Office was included in the later Windows XP system sales pitch (Whitworth, 2005). Microsoft failed to realize how Clippy was perceived. “I HATED that clip. It hung around watching you with that nasty smirk. It wouldn’t go away when you wanted it to. It interrupted rudely and broke your train of thought. It never actually had an answer to questions I had.” (Whitworth, 2005). Microsoft spent 25,000 hours testing the technical capabilities of Clippy, but ignored the social components critical to ensuring a rich coupling between Clippy and the human user, dooming Clippy to failure.

Thus, in addition to considering the technical aspects related to the collaboration between humans and automation, it is also necessary to consider the social aspects. What does it mean for an automaton to be an effective team player? How does an automaton earn an operator’s trust? How is it possible to strengthen the bonds among human and autonomous teammates? How can an automaton assist, without interrupting human processes or undermining human capabilities?

\textsuperscript{6} “Clippy” is a nickname; “Clippit” is its official name.
A Collaborative Systems Approach: Complementing Capabilities

“Basically, meaningful interaction with an environment depends upon the existence of a set of invariate constraints in the relationships among events in the environment and between human actions and their effects” (Rasmussen, 1983).

If there is a rich coupling between the components outlined in Figure 1, the human-machine team will jointly bridge Hutchins, Hollan, and Norman’s Gulf of Evaluation and Gulf of Execution (Hutchins, Hollan, & Norman, 1985; see Figure 2A). However, if there is a poor coupling between the humans and technologies, then another gulf is introduced creating additional uncertainties for each component (see Figure 2B).
Figure 2. The Gulfs of Execution and Evaluation (adapted from Hutchins, Hollan, and Norman, 1985).

On one side of these gulfs resides the human-machine team, with their goals and intentions. On the other side is the work domain. The size of the gulf of execution depends on the effectiveness of the actions or controllability the human-machine team has to achieve his or her goals. The size of the gulf of evaluation depends on how well the human-machine team can observe, perceive, and understand the state of the world in terms of their intentions (e.g., see the best move on the chess board). The goal-directed interaction between the human-machine team and the physical system is dependent upon
constraints. To span the gulf of evaluation it is vital to make the constraints of the work
domain salient to the human-machine team. To span the gulf of execution the constraints
associated with the human-machine team — designing controls and displays that are
consistent with a human operator’s reasoning capabilities — must be considered.
Rasmussen’s Skills, Rules, Knowledge (SRK) framework defines three ways in which
constraints might be represented — signals, signs, and symbols — which, in turn,
distinguish three levels of human performance — skill-based, rule-based, and knowledge-
based.

_Skill-based_ behavior consists of sensory-motor tasks without conscious control
(Rasmussen, 1983). In _rule-based_ behavior, an individual has a set of predetermined
solutions that are triggered by specific conditions, or signs. _Knowledge-based_ behaviors
occur when an individual encounters a novel unexpected situation for which no procedure
exists (Rasmussen, 1983).

Consider a chef preparing a meal — chopping vegetables or continuously
adjusting the gas flame of a burner to perfectly fry an egg are skill-based behaviors, and
following a recipe is rule-based behavior. Now imagine that the recipe calls for vanilla
extract — but when the chef looks in the pantry, the bottle is empty. The chef considers
his or her options: (1) leave the cooking process, which is currently in a critical stage to
acquire more vanilla extract; (2) skip the vanilla extract; or (3) find a suitable substitute.
When the chef decides to try almond extract as a substitute, he or she is exhibiting knowledge-based behavior.\(^7\)

The prosthetic/substitution approach attempts to replace the user at all three modes of interaction. While some researchers have suggested that autonomous systems are easiest to develop for skill-based behaviors and hardest for knowledge-based behaviors (Cummings, 2014), building autonomous systems that replace humans can be very challenging for all three categories. For example, cutting vegetables is simple for an experienced sous chef, but very difficult for an autonomous system (Lenz, Knepper, & Saxena, 2015). Rather than replace the human, the focus should be on designing autonomous systems that support the human in all three modes of interaction — skill, rule, and knowledge-based.

Autonomous systems can be designed to augment a human’s SRK behaviors in several ways. First, the autonomous system should help the human explore the state space. IBM is currently repurposing Watson to help chefs do this very thing. Imagine a chef that has just returned from the garden with a plentiful bounty of sweet corn, lima beans, zucchini, and onions but is at a loss for what to do with them. Using the Chef Watson app the chef can enter these ingredients and get back a variety of recipes.

\(^7\) Note that if the almond extract replacement is deemed a success, the chef will likely switch to rule-based behavior in a similar situation in the future – “I am out of vanilla extract; I will use almond extract instead.”
featuring these ingredients — including zucchini tacos, zucchini fricassee and zucchini curry.

The chef can then collaborate with Chef Watson by adding constraints and narrowing the search space. For example, if the chef is in the mood for a specific type of cuisine, he or she can select the “Pick a Style” option. By default, Watson will recommend styles based on the ingredients — in this case Watson suggested Peruvian, Basque, Nuevo Latino, Moroccan, Tailgating, Tex Mex, Nashville, Cajun, and Israeli — but also provides the option for the chef to select an out of the box cuisine for these ingredients, like Japanese. Table 2 illustrates how Chef Watson has modified the zucchini taco recipe to infuse the dish with Japanese flavors. For example, Manchego cheese, which has a flavor profile similar to miso, has replaced goat cheese.

Table 2. Recipe Differences highlighted in gray.

<table>
<thead>
<tr>
<th>Zucchini Taco</th>
<th>Japanese Zucchini Taco</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egg</td>
<td>Egg</td>
</tr>
<tr>
<td>Lima Bean</td>
<td>Lima Bean</td>
</tr>
<tr>
<td>Onion</td>
<td>Onion</td>
</tr>
<tr>
<td>Corn</td>
<td>Corn</td>
</tr>
<tr>
<td>Zucchini</td>
<td>Zucchini</td>
</tr>
<tr>
<td>Vegetable Oil</td>
<td>Vegetable Oil</td>
</tr>
<tr>
<td>Butter</td>
<td>Butter</td>
</tr>
<tr>
<td>Thai Chile</td>
<td>Chile de Arbol</td>
</tr>
<tr>
<td>Lemon Grass</td>
<td>Jalapeno Pepper</td>
</tr>
<tr>
<td>Pineapple Juice</td>
<td>Lemon Juice</td>
</tr>
<tr>
<td>Goat Cheese</td>
<td>Manchego Cheese</td>
</tr>
<tr>
<td>Flour Tortilla</td>
<td>Flour Tortilla</td>
</tr>
<tr>
<td>Coriander Seed</td>
<td>Caraway Seed</td>
</tr>
</tbody>
</table>
At this point, should the chef decide to make the Japanese zucchini tacos, Chef Watson has a recipe queued up and ready for the chef to access. And if the chef heads to the pantry and can’t find any caraway seed, luckily Chef Watson has it covered — potential substitutes are generated upon request. The chef, aided by his or her sous chef, Chef Watson, can now get busy cooking.

The Chef Watson app illustrates how an automaton can be used both to complement cooking skills, and potentially stimulate a chef’s creative ability to invent new solutions to the cooking problem. Note that the point here is neither to use Chef Watson as a substitute for the human, nor to use it to enforce procedural compliance on the human. Rather, Chef Watson becomes a creative partner — it helps the ‘team’ explore the cooking problem efficiently and effectively, think productively, and experiment with innovative alternatives.

**A Collaborative Systems Approach: Creating Social Cohesion**

Developers must also consider social aspects in order to facilitate a rich-coupling between the human and autonomy. In 2004 Valve Corporation released *Half-Life 2* (*HL2*), the successor to their massively successful 1998 debut *Half-Life*. *HL2* is a first person shooter game in which gamers play as Gordon Freeman, a scientist who finds himself inspiring a resistance movement against a conquering alien force in a dystopian future. The *HL2* series deftly filled the shoes of its beloved predecessor by developing several innovative gameplay elements, including a fully realized AI sidekick named Alyx.
Vance. Alyx received almost universal praise, often ranking at the top or near the top of lists ranking the greatest non-playable characters of all-time, even years after her introduction (Martin, 2008; Dodd, n.d.).

Alyx’s success is a credit to Valve’s intense development and play testing process. Over 100 actresses were auditioned to provide Alyx’s voice, with developers seeking a voice actress that could be charming and warmly intimate, but could also be strong, confident, and believable (Hodgson, 2004). During play testing it became clear that having Alyx be capable of providing assistance to the player wasn’t enough. Developers initially tried to create a sense of urgency by having Alyx say things like “Hurry up!” and “Keep Moving!”; however, players felt liked they were being nagged, and ended up hating Alyx. This led to a major design change — having Alyx almost always follow the player, rather than leading the way (Wood, 2004).

Ultimately, making Alyx likeable was just as important — if not more so — than making her capable. The developers designed multiple scenes to humanize and endear Alyx to the player, and each scene had multiple variables that had to be just right.

If you don’t like Alyx, you’re not going to have much fun with Episode 1. So Alyx being likeable was one of our most crucial design goals. Little moments like the

---

8 Unfortunately, needing to be “likeable in addition to being capable” is something many women experience all the time. The role that the perceived gender of an AI plays is a very interesting topic, yet is outside the scope of this effort.
Zombine joke are designed to make Alyx more endearing. Surprisingly, lighting was really important too. Under red light, Alyx’s self-deprecating groan looked more like she was sneering at the player for not getting the joke. Changing the lighting to blue and then adjusting the direction of the light so that it changed the shadows on her face fixed the problem. (Wolpaw, 2004).

By the time the player reaches the end of HL2: Episode 2 they have spent many hours working with Alyx towards a common goal, and most players will have developed an emotional attachment to her, which Valve uses to devastating effect. At the end of HL2: Episode 2 the player watches helplessly as Alyx’s father is brutally murdered right in front of her. As Alyx — who is also restrained — screams in rage and agony, there’s an incredibly brief moment when she glances back at the player, whispering “Gordon”, her eyes pleading with the player for help. This gut-wrenching sequence continues as Alyx clings to her father’s lifeless body, and her desperate sobs remain even after the screen fades to black. This is how developers can create an AI that connects with a user on an emotional level.

Everything Valve got right with Alyx Vance, Microsoft got wrong with Clippy. Valve realized during the design process that Alyx should follow the player’s lead and not control the action. Clippy would show up uninvited, take control of the user’s mouse

---

9 “Zombine” is a portmanteau of zombie and combine. The main antagonist force in Half-Life 2 is known as The Combine.
cursor, and keep coming back no matter how many times the user sent it away (Whitworth, 2005). While Valve discovered that in the wrong lighting, what was intended to be a humanizing groan could be perceived as looking down on the player, Clippy’s tone always seemed to convey that it knew better than the user. Valve spent thousands of hours perfecting Alyx’s interaction with the player, resulting in one of the most beloved video game characters ever; Microsoft did not — and ended up with perhaps the most notorious automated assistant ever.

**Summary**

As many researchers have noted repeatedly (e.g., Klein et al., 2004; Dekker & Woods, 2002), effectiveness in sociotechnical systems will often depend on whether the technologies function as collaborative systems or team players. In many complex work domains, success is beyond the capabilities of un-aided humans, yet human capabilities are often critical to ultimate success. An important motivation behind the Cognitive Systems Engineering approach was the realization that no matter how carefully designed, all automated control systems will eventually face situations that were not anticipated at the time of their design (Flach, 2015). Thus, at some point the human operators of those systems will be called upon to complete the design. In other words, the human operators will need to intervene to creatively deal with the requisite variety that was not anticipated by the designers of the automated systems. The SRK framework is specifically geared towards drawing attention to user interfaces, and ways to design representations so that
the human and automated systems can work together to creatively respond to the inevitable, unanticipated variability endemic to complex work domains.

In sum, the challenge is to move beyond an either/or attitude with respect to humans and technology — the classic “Humans Are Better At, Machines Are Better At” lists — that tends to focus on optimization of the separate human and autonomous components as the top priority, and leaves the design of interfaces and team processes as an afterthought. The alternative is to take a holistic perspective and to begin thinking in terms of both/and, where the goal of design is a seamless integration of human and technological capabilities into a well-functioning sociotechnical system. Success in complex domains will ultimately depend on the ability of humans AND technologies working together as well coordinated teammates — each contributing unique abilities to create a team with the potential to be greater than the sum of its parts, and thus jointly bridge the gulfs of execution and evaluation in order to address the requisite variety of complex domains, or wicked problems.

The goal of this dissertation is to develop a sociotechnical system that integrates human and autonomous capabilities to enable a single operator to effectively manage multiple UVs. Papautsky, Dominguez, Strouse, and Moon (2015) identify three stages for successfully integrating autonomous systems into UV operations: (1) Understand — develop an accurate representation of the domain and ensure that the autonomous system can reason about critical domain aspects; (2) Generate — design interfaces that enable a
rich coupling between the human operator and the autonomous system; and (3) Validate — evaluate the human-autonomy team in a realistic test environment. Each of these stages is described in-depth in the ensuing chapters.
CHAPTER II. Understand

Historically, the primary focus of psychological research has been on understanding and modeling what is inside a person’s head and using these individual characteristics to attempt to predict performance across a wide range of situations (Behymer, Mateo, & McCloskey, 2015). As a result, researchers not only lack adequate descriptions of the environment, but also fail to recognize the need for such a description despite the overwhelming evidence that context matters (Heft, 2001). Ecological psychology, which seeks to explain human behavior in the context of the environment in which it is occurring, has attempted to address this deficit.

The roots of ecological psychology lie in the links between the social psychologist Roger Barker, perception psychologist J.J. Gibson, and the radical empiricist position of William James who all argued that psychological processes can only be fully understood in terms of pragmatic adaptations to ‘situations.’ (Heft, 2001). For example, Barker (1968) suggested that one could more accurately predict a child’s behavior from knowledge of the situation the child was in — in class, at a sporting event, or hanging out with friend — than from the knowledge of the behavioral tendencies of a particular child — whether the child is shy or outgoing. Therefore, research to develop autonomous tools to support operators in complex mission environments must begin with an understanding of the work domain. Researchers and developers can gain this understanding using cognitive task analysis (CTA).
Cognitive Task Analysis

Cognitive Task Analysis is a set of methods used to identify the cognitively demanding tasks operators face as well as the cognitive skills necessary to perform these tasks proficiently (Militello & Hutton, 1998). The term CTA first appeared in the late 1970s and emerged as a way to understand how situational dynamics impact performance (Militello & Hoffman, 2008). One of the key insights of this research was that experts did not consider the pros and cons of various plans in time critical situations; rather, they often considered a single option after recognizing key situational elements (Klein, Calderwood, & Clinton-Cirocco, 1986). Fire fighters were not so much making decisions; rather, they were engaged in an ongoing process of modifying their actions to meet situational demands (Klein, 1989). Many CTA methods have been developed to capture the situational elements experts use to identify a course of action.

Crandall and Calderwood (1989) for example, used the critical decision method to study decision making in neonatal intensive care units (NICU). They found that experienced NICU nurses were often able to identify premature infants who had developed sepsis.\(^\text{10}\) However, when the NICU nurses were asked how they knew an infant had sepsis, many initially responded by saying, “I had a gut feeling” or “I just had an intuition”. Clausewitz\(^\text{11}\) described intuition as the flash of insight when an individual

---

\(^\text{10}\) Sepsis occurs when the immune system attacks the body’s own organs and tissues in response to an infection.

\(^\text{11}\) Carl von Clausewitz, a Prussian general who wrote the classic military theory *Vom Kriege* (On War). For an excellent treatise on Clausewitz see Jon Sumida’s *Decoding Clausewitz: A New Approach to On War*. 
knows the right course of action despite the presence of uncertainties (as cited by Duggan, 2005). Klein (1998) defined intuition as the “ability to use experience to recognize the key patterns that indicate the dynamics of the situation”. CTA methods are designed to unpack and demystify intuition by eliciting these key patterns from experienced operators.

In the critical-decision method, an interviewer begins by asking an expert to think of a time when his or her skills were really challenged — “Can you think of a time when it was really difficult to diagnose that an infant had sepsis?” Once the incident is identified, the interviewer asks the expert to provide a quick overview of the incident, noting when the expert uses phrases like “I just had a feeling”. After the incident’s timeline has been verified, the interviewer rehashes the incident and focuses on unpacking those “gut feelings” by asking questions like, “What were you noticing when you started to think the infant might have sepsis?” in order to identify the critical cues the expert used to inform his or her decision making. Finally, the interviewer uses “What If” queries like “How would this situation had played out if this had been your first day on the job?” to identify common errors a novice might make in the same situation.

The critical-decision method allowed Crandall and Calderwood to identify a list of diagnostic cues that NICU nurses were using to identify sepsis in premature infants. For instance, irritability is a symptom of sepsis in adults; however, premature infants actually become less irritable with sepsis. Capturing these cues is vital for developing
assessment techniques for selecting personnel, training existing personnel, and improving
decision-making support tools like IMPACT’s CCAs, IAs, and autonomic frameworks.
The remainder of this chapter describes how CTA methodologies were used to examine
UV and base defense operations in order to inform the design of IMPACT’s autonomous
tools.

*The Domain: Base Defense*

Base defense was chosen as the problem domain for several reasons. First, base
defense is a relevant mission for multiple services, and a notional base set on the coast
provided a realistic and challenging scenario for involving UAVs, UGVs, and USVs.
Second, integrated base defense strategies have shifted from reactive — responding to
threats — to proactive operations focused on intelligence-driven targeting, thus
supporting a wide range of possible events and mission complexity. Therefore, an
understanding of the base defense domain was required to inform the design of the
autonomous tools, the HAI, and the evaluation scenario. Unfortunately, experienced
operators who manage multiple heterogeneous UVs in support of base defense missions
do not exist. To overcome this limitation, CTAs were conducted with several types of
participants, with each type articulating different slices of the problem space, including
search and rescue, base defense, and UV operations.
Cognitive Task Analysis: Search and Rescue

The goal of this CTA was twofold: (1) understand how search and rescue personnel choose search patterns in order to inform the development of an IA that could recommend search patterns; and (2) understand how search and rescue personnel determine which search vehicle to use in order to inform the development of an IA that could recommend vehicles. One participant, a male Civil Air Patrol (CAP) captain with 14 years of coordinating air and ground search efforts using manned air and ground vehicles in emergency response and wilderness search operations was interviewed. Two CTA methodologies were used.

First, the critical incident method (Klein, Calderwood, & MacGregor, 1989) was used to identify situations in which the participant conducted a cognitively challenging search and rescue mission. Second, a simulation interview (Militello, Hutton, Pliske, Knight, & Klein, 1997) was used to ask the participant how he would determine the search pattern and search vehicle to use in a variety of situations such as a man overboard, searching for a high value target, or searching for an unidentified watercraft. The participant’s responses provided insight on choosing the best search pattern in a given situation (see Figure 3) as well as choosing the best search vehicle to accomplish mission objectives.
Search Pattern Selection

Figure 3 shows the decision space for determining the best search pattern in a given situation. For example, if a search and rescue professional received a task to search for a man overboard, he or she would use a route (track line or track crawl) search because the search area isn’t mountainous, the ship’s route is known, and the man overboard is likely somewhere along the route.

Figure 3. Optimal Search Pattern Decision Tree.

Vehicle Selection

At a high level, when making a decision about which search vehicle to use, the participant considered the following in order of priority:
1. *Does the vehicle have the capability to complete the task?* At the most basic level, can the vehicle accomplish the task? This includes determining if the vehicle has the right sensor (e.g., infrared [IR] if searching at night), if the vehicle has enough fuel to complete the task and return to base, if the vehicle can remain in communication range while completing the task, and if the vehicle can handle current weather conditions. It also includes considering if the vehicle can handle likely future developments such as taking kinetic action against an enemy or rescuing the search target.

2. *How quickly and efficiently can the vehicle complete the task?* In most cases, the faster the better, but the impact of fuel consumption — an air vehicle could get there faster, but a ground vehicle would minimize fuel usage — is also considered.

3. *What impact does assigning the vehicle to the task have on current and potential future tasks?* For example, if one UAV has the ability to act as a communications relay, it might be best to keep that UAV unassigned in case a need for a communications relay arises.

4. *What impact does assigning the vehicle to the task have on its maintenance schedule?* UVs have a maintenance schedule, and every second of operating time brings the UV closer to mandatory servicing and mission downtime.
This domain knowledge was used to develop a cognitive domain ontology (CDO) — a representation of domain knowledge that IAs use to categorize situations, develop hypothesis, and plan and recommend courses of action (Atahary, Taha, Douglass, & Webber, 2015). Within IMPACT, the CDO allows the IA to sort various courses of action by mission relevant variables. For example, in a given situation, the IA could use the CDO to determine the UV most likely to find a target, the UV that could reach the search area in the least amount of time, and the UV that could reach the search area using the least fuel.

**Cognitive Task Analysis: Base Defense**

In order to gain an understanding of current base defense operations as well as how UVs could be integrated into base defense operations, a CTA was conducted with Security Force Squadron personnel currently assigned to the Base Defense Operations Center (BDOC) at Wright-Patterson Air Force Base. Five participants — four male, one female — from the United States Air Force Material Command 88th Security Forces Squadron were interviewed. Each participant had at least one deployment experience conducting security operations, with locations including Afghanistan, Iraq, Kuwait, Saudi Arabia, and Turkey.

First, the task diagram method (Militello, Hutton, Pliske, Knight, & Klein, 1997) was used to elicit the major job components of base defense personnel and identify the major cognitive challenges base defense personnel face. Next, the critical decision
method (Klein, Calderwood, & MacGregor, 1989) was used to identify critical incidents security forces had experienced in order to inform evaluation scenario development. Finally, the simulation interview method (Militello, Hutton, Pliske, Knight, & Klein, 1997) was used to investigate how security force personnel would use UVs in the context of a notional base defense scenario.

Using PowerPoint slides, participants were shown a mock base and asked how they would defend it using their current methods and equipment. Participants were then briefed on the capabilities of a UAV, UGV, and a USV — Boeing’s ScanEagle, General Dynamics’ Mobile Detection Assessment and Response System (MDARS), and BAE System’s Protector respectively — and asked how they would incorporate these UVs into the mock base’s defense operations. Participants were then given a series of simulated events — a crowd forming outside a base gate, a HUMINT (human intelligence) report of a suspicious vehicle — and asked how they would respond using these UVs.

This CTA provided insight into both the constraints of the base defense domain as well as the procedures that security force personnel use to respond to specific threats. In addition to the constraints identified in the CTA with search and rescue personnel (capability, time, impact on other tasks, and maintenance) security personnel identified additional constraints that need to be considered including detectability (the sound intensity of the UV’s engines as well as the UV’s size), presence (the size and weaponry of the UVs), tracking (a UV’s ability to track targets), and crowd control (a UV’s ability
to use non-lethal methods to control crowds). These constraints, along with environmental constraints (e.g., how the UV’s perform in specific weather) and the size of the target were incorporated into the IA’s CDO, allowing the IA to compare and rank potential plans in response to operator play calls (Hansen, Calhoun, Douglass, & Evans, 2016).

Additionally, the critical incidents interviewees discussed provided a list of real-world events (e.g., suspicious vehicles, mortar attacks, etc.) that base defense personnel must respond to as well each event’s Quick Reaction Checklist (QRC) — a list of actions to take in response to potential security threats. This list of real-world events was used to create realistic experimental scenarios in which to evaluate IMPACT (see Chapter V for further detail). The simulation interview results also helped determine how base defense personnel would use UVs to respond to each type of security threat. These results informed the development of QRCs for UV operators tasked with supporting base defense missions.

**Cognitive Task Analysis: UAV and UGV Operations**

The final CTA was conducted with experienced UAV and UGV operators including one retired Air Force pilot with extensive experience using UAVs in support of base defense operations and five Army UGV operators from the U.S. Army Maneuver Center of Excellence at Ft. Benning. This set of interviews investigated if UVs could realistically conduct the tactics and techniques identified during our interviews with the
CAP Captain and BDOC personnel. For example, during interviews with BDOC personnel, a participant suggested that a UGV could be used to patrol along the shoreline. However, several UGV operators stated that the shore’s terrain may be too unstable for certain types of UGVs.

**Cognitive Task Analysis: Informing Autonomous Capabilities**

In this section, an example—security force personnel responding to a suspicious vehicle—will be used to illustrate how the CTA results informed the development of IMPACT’s autonomous capabilities. Security force personnel indicated that when a suspicious vehicle is reported, a manned Quick Reaction Force responds to the threat while other manned patrol units change their tactics to reflect the heightened state of alert. Security force personnel further indicated that if a UV capability was available it would be beneficial to get surveillance on the suspicious vehicle as soon as possible. Interviews with search and rescue personnel also identified the importance of ensuring that the UV would be capable of detecting the suspicious vehicle given current environmental conditions, as well as considering the UV’s fuel usage and current tasking. Interviews with UV operators provided additional information about UV route planning (e.g., the importance of avoiding no-fly zones), sensor control (e.g., keeping the target within the sensor’s field of view), monitoring the ongoing task to ensure everything was on plan (e.g., will the UV arrive on time?), and observing the sensor feed. Figure 4 outlines this process from start to finish.
By capturing this information, IMPACT’s autonomous systems were tailored to support the operator. For example, by creating a CDO that represents the constraints (probability of finding the target, time enroute, and fuel efficiency, as well as the impact of changing the UV’s current tasking), IMPACT’s IAs can rank and compare plans to determine which UV is the best to use. CCAs can rapidly calculate the most efficient route for a vehicle to take to a specific point while taking into account no-fly zones, unpassable terrain, and environmental conditions (Kingston, Rasmussen, & Mears, 2009). Finally, the Rainbow autonometrics framework (Verbancsics and Lange, 2013) can monitor the situation and alert the operator when a deviation from the plan occurs, such as a
strong headwind delaying a UAV. With these autonomous capabilities, a multi-UAV operator can define a high-level task (e.g., find the suspicious vehicle) and focus on observing the sensor feed (see Figure 5).

![Decision Process for Suspicious Vehicle with Autonomy](image)

Figure 5. Decision Process for Suspicious Vehicle with Autonomy.

**Summary**

CTA methods were used to identify the critical situational factors experienced search and rescue workers, security force personnel, and UV operators use to achieve mission effectiveness in base defense operations. The results of the CTA informed the development of autonomous capabilities that could plan UV routes (CCA), recommend
the best strategy and UV (IA), and identify when a plan has gone south (Rainbow Autonomics framework). Though these autonomous capabilities have the potential to improve the abilities of the multi-UV operator, a HAI that allows the operator to communicate his or her intended goals to the autonomous system and to understand autonomous system states was needed. Bridges that span the gulf of execution and the gulf of evaluation discussed in Chapter I must be designed and built to enable effective human-autonomy collaboration.
CHAPTER III. Generate

“Several years ago Rasmussen (personal communication, December 7, 1999) kindly reviewed some process control graphical displays of ours and remarked that ‘display designs just seem to appear from the blue sky like works of art. I think it would be very productive and influential if you could describe how you selected the visual designs’” (Bennett & Flach, 2011)

In this chapter, HAIs that bridge the Gulfs of Execution and Evaluation described in Chapter I and that enable a single operator to collaborate with autonomous tools to manage multiple heterogeneous UVs in base defense operations will be described. The science behind these interfaces, including Ecological Interface Design (EID) theory and display design principles such as direct perception and direct manipulation will be brought to the fore. Empirical evaluations of early display prototypes will be presented and described. However, the essential role that serendipity and art played in the design process will also be acknowledged. In fact, art’s role in the design process is often undervalued and unexamined within the human factors literature. To remedy this situation, and to shine a light on the black box that is often the creative design process, the following section — with a nod to Wylie (1951) — is an essay on the art of design, extraneous to the chapter and yet its theme, which the impatient may skip and the reflective might enjoy.
The Art of Design

“You are here to learn the subtle science and exact art of potion-making. As there is little foolish wand-waving here, many of you will hardly believe this is magic. I don't expect you will really understand the beauty of the softly simmering cauldron with its shimmering fumes, the delicate power of liquids that creep through human veins, bewitching the mind, ensnaring the senses. . . I can teach you how to bottle fame, brew glory, even stopper death — if you aren't as big a bunch of dunderheads as I usually have to teach.” (Rowling, 1997).

In a nod to the above quote from Harry Potter and the Philosopher’s Stone, Flach and Bennett (2011) titled their book on display design Display and Interface Design: Subtle Science, Exact Art. This title perfectly captures the two sides of design, art and science and rightly gives equal weighting to both. However, many human factors professionals would claim that science alone is sufficient for display design, that human factors theories, principals, and design guidelines are the necessary precursors to a successful design, and that theory driven design is the optimal solution to a design problem. This attitude is often pervasive within academic communities and isn’t limited to human factors. Pirsig (1991) stated:

“You can imagine the ridiculousness of an art historian taking his students to museums, having them write a thesis on some historical or technical aspect of what they see there, and after a few years of this giving them degrees that say they
are accomplished artists. They’ve never held a brush or a mallet and chisel in their hands. All they know is art history.”

To understand why this attitude is misguided, it’s important to examine how display design principles and guidelines are generated in the first place. Display design theory development is not a prospective forward-looking phenomenon; rather it is a retrospective, backward-looking one. For example, after the Three Mile Island accident, it was clear that one of the precipitating causes was poorly designed control panels. However, this only became obvious after the incident. As Weick (1995) states, if the outcome is considered poor, the actions leading up to the outcome are reconstructed to emphasize the errors, flaws, and inaccuracies of these actions, even if these errors were not obvious at the time. Unfortunately, this implies that errors should always be anticipated and that good perceptions, analyses, and discussions will always lead to good results (Starbuck and Milliken, 1988).

Design, as juxtaposed with design theory, is a forward looking phenomenon. In fact, designing innovative interfaces is, in many ways, an artistic endeavor, and as Rasmussen observed, do seem to appear out of the blue sky. As Solnit (2014) states, the methods for accomplishing creative work are always unpredictable and cannot be reduced to replicable formulas. Unfortunately, as Flach and Voorhorst (2016) point out, when it comes to design, people are too often searching for a process to follow or a list to check.
The distinction between design theory and design can best be illustrated through an analogy. In elementary school there are usually two types of writing classes, creative writing and literature. Typically, creative writing classes are phased out as students proceed through the school system, with more and more emphasis placed upon the rules and regulations of formal writing. Students are provided a formula — how to make note cards, how to construct an outline, and how to support each major point with at least three items. While this procedure may certainly aid in authoring a literary criticism of a novel, this approach will never produce a novel! The methods used to teach formal writing are inadequate for teaching creativity.

Similarly, the way design theory is taught — providing students with lists of display design principles — is not conducive to teaching students how to design. A good design, like a good work of fiction, will never be the result of a top-down procedure that starts with step one and outputs a brilliant design at step fifty. How then should design be learned? If design is more akin to writing a novel than writing literary criticism, how should design students proceed? An old joke provides some guidance:

A tourist is wandering around New York City and is clearly lost. He walks up to a local and asks, “How do you get to Carnegie Hall?”. The local responds, “Practice, man, practice”.
Magnus Carlsen — briefly mentioned in Chapter I as obtaining the highest Elo rating of any human chess player — became history’s youngest world No. 1 chess player at the age of 18. A 2010 profile in *Time* quoted Garry Kasparov as saying, “[Carlsen] has a natural feel for where to place the pieces… Before he is done Carlsen will have changed our ancient game considerably” and experts watching Carlsen’s games are often surprised by his moves, only later realizing his choices were perfect (Harrel, 2010). Even Carlsen himself has a difficult time describing his ability — “It’s hard to explain, sometimes a move just feels right.” (Harrel, 2010). The *Time* profile portrays Carlsen’s chess prowess and innovative playstyle as an innate ability, the origins of which are an unsolvable mystery.

By most accounts Kobe Bryant was one of the best basketball players in the history of the National Basketball Association (NBA); he ended his career with five championship rings, 33,643 points (3rd all time), 11 first team all-NBA selections, and 18 all-star appearances. Bryant’s success is often attributed to “god-given ability” and of course, it helps to be 6’6. However, if size and athletic ability were all that mattered, as Ballard (2008) put it, Eddie Curry would be all-NBA and Derrick Coleman would be getting ready for his hall of fame induction ceremony. A key to Bryant’s success was

---

12 Eddie Curry, an NBA journeyman, was considered to be among the best high school players in the United States his senior year and was drafted 4th overall in the 2001 NBA draft. Derrick Coleman was taken 1st overall in the 1990 NBA draft and was expected to have a career similar to Hall of Famers Karl
practice. Every day he made 700 to 1,000 shots, in addition to 4 hours of weight trainings and conditioning (Men’s Fitness, n.d.). Bryant’s method was consistency — “You have a program, and a schedule, and you have to abide by that, religiously. You just stick to it, and it's the consistency that pays off.” (Men’s Fitness, n.d.). Additionally, at the end of each season, Bryant sat down with his coaches to break down the season and establish goals and an improvement plan for the off-season. For example, one year while many of his peers were relaxing on the beach or focused on refining existing skills, Bryant worked with Hall of Famer center Hakeem Olajuwon to improve his post-up moves, adding yet another skill to his repertoire.

Ericsson, Krampe, and Tesch-Romer (1993) suggested that a key to achieving mastery in a specific area is the amount of deliberate practice an individual performs. Though it is easy to attribute Bryant’s success to his “god-given talent”, the amount of hours he put in at the gym are just as crucial to his success. Similarly, the genius and creativity Magnus Carlsen displays while playing chess are often attributed to mysterious factors such as intuition or innate talent. However, a closer look at Carlsen’s daily routine sheds light on his talent’s origin. Lehrer (2010) described how living in the computer chess age has allowed Carlsen to play multiple games at once against sophisticated chess playing algorithms, giving him an unprecedented amount of deliberate practice. While the amount of deliberate practice previous generations of Malone and Charles Barkley. Coleman appeared in a single All-Star game and is considered to have never reached his full potential.
chess players could get was limited by the number and stamina of quality opponents they could find, Carlsen played more high-quality games by the age of 13 than many grandmasters had their entire lives. This deliberate practice provided Carlsen with so much experience that he is able to utilize the knowledge he gained through it at a level so automatic it appears intuitive.

Thus, deliberate practice is key when you want to improve a skill. This seems like an obvious point, and most people grasp this idea when considering an activity like driving a car or riding a bicycle. However, there seems to be a block when it comes to people like Carlsen and Bryant, whose talents are often credited to innate ability and the importance of practice is overlooked. Why is this? Perhaps, if skill is due to innate talent it lets everyone else off the hook. Considering the role that deliberate practice plays may place the blame too close to home. As Ericsson, Krampe, and Tesch-Romer (1993) state, there is nothing fun about deliberate practice. In fact, the desire to practice, even though deliberate practice is not fun, is another characteristic vital for skill development:

“There’s a difference between loving basketball and liking basketball. There are only about 30 guys in the league who love it, who play year round. Allen Iverson loves to play when the lights come on. Kobe loves doing the shit before the lights comes on. This thing, this freakish compulsion, may be the hardest element of the game to quantify. There are no plus-minus stats to measure a player’s ruthlessness, his desire to beat his opponent so badly he’ll need therapy to recover.” (Ballard, 2008).
Carlsen has a similar attitude towards chess; when asked if he saw chess as a game of combat or a game of art Carlsen replied, “Combat. I am trying to beat the guy sitting across from me and trying to choose the moves that are most unpleasant for him and his style. Of course some really beautiful games feel like they are art, but that's not my goal.” (Harrel, 2010).

The will to succeed and win is so prominent in both Bryant and Carlsen that they are able to overcome the negatives associated with deliberate practice. In his memoir/writing tutorial *On Writing*, Stephen King (2000) related a story about his son’s desire to take saxophone lessons. King bought the instrument and his son started taking lessons. Six months later King talked to his son about quitting and his son agreed. King knew his son wasn’t interested in being a saxophone player, neither because his son didn’t attend practice — he did regularly — nor because his son wasn’t technically adept at it — he appeared to know all the notes. The problem was King’s son only practiced during designated practice times — there was no creative discovering on his own. King’s son never took out the saxophone on his own time because he enjoyed it. King ended the anecdote by stating that to achieve greatness, “you do it (whatever it is) until your fingers bleed or your eyes are ready to fall out of your head”.

Both the amount of deliberate practice and the willingness to engage in practice are critical to skill development. However, there is one additional critical factor: type of practice. Bill Belichick — head coach of the New England Patriots — has won four
Super Bowls and was a miracle play away from the first 19-0 season in the history of the National Football League (NFL). What separated Belichick’s Patriot teams from the rest of the NFL? According to Gasper (2008) the answer is situated practice. Gasper interviewed a former player who described his first day of training camp under Belichick. Belichick created the following situation during the practice: the Patriots were losing by a field goal, had the ball at their own 17 yard line with 1:21 left in the game, and zero timeouts — an exact recreation of the situation the Patriots faced when they went on to win Super Bowl XXXVI. Belichick had his players practice scenarios like this during training camp so that by the time they needed to perform perfectly in the post-season they would be ready. Belichick understood that in addition to drills and conditioning, players needed to practice under game-specific constraints.

Constraints are essential for deliberate practice, as Edward Zagorski—a design professor who spent over thirty years teaching industrial design at the University of Illinois and originator of the egg drop problem—discovered. Without constraints, Zagorski found that students tended to be overwhelmed by the infinite possibilities; the freedom to do anything actually hinders creativity and innovation (Murray-Tiedge, 2007). For example, Zagorski initially gave his students a block of wood and instructed them to cut it and reconfigure it into a new form; the results were unoriginal, bland, and tired designs (Zagorski, 2011). Zagorski then added a key constraint—students were required to make at least one cut, but could not make more than three cuts. Suddenly, the students were creating unique one-of-a-kind solutions that neither the student nor the professor
envisioned when the task was assigned (Murray-Tiedge, 2007). Zagorski’s three-cut problem provided enough constraints to allow the student to focus, but not so many that they were over constrained.

In sum, any skill, including design, can be improved through deliberate practice, as long as the individual is able to commit to a significant amount of deliberate practice and that deliberate practice has been tailored to hone the critical skills needed to succeed at the target task. Stephen King recommended that would-be writers should write a lot and read a lot. The same advice applies to would-be designers: design as much as possible and get exposed to as many different types of designs as possible. Exposure to different design ideas is key, because designers often find solutions for problems they are working on in the design concepts of others. In fact, good designers know that inspiration can hit at any time, and are prepared for it.

Preparing for Serendipity

In 2000, Douglas Caldwell—an employee working in the Topographical Engineering Center of the United States Army Engineer Research and Development Center—had a problem. In fact, Caldwell’s problem had been vexing military leaders for thousands of years—how to create 3D terrain models of the battlefield (Shedroff & Noessel, 2012). Standard practice involved creating a model using sand tables, which were not reconfigurable and difficult to transport. On a night out at the movies, as Caldwell and his son were watching X-Men, Caldwell was stunned to find a better
method on the big screen (Shedroff & Noessel, 2012). During the movie, mission tactics were planned using a 3D reconfigurable table topographical map. Caldwell was so inspired by this scene that he immediately solicited proposals to create a reconfigurable dynamic sand table that could create a physical terrain model based on digital terrain data (United States Army, n.d.). Four years later, the XenoVision Mark III was completed. (Francica, n.d.)

Archimedes’ bathtub, Newton’s apple, Goodyear’s hot stove, and Caldwell’s movie, exemplify Solnit’s (2014) claim that it is actually distraction that moves imagination forward, not uninterrupted focused concentration. However, designers must be prepared, so that when the Eureka moment arrives, they can capitalize on it. Johnson (2012) stated that most good ideas come to people as hunches, which may sit in the back of a person’s mind for months or years. The problem with these hunches is that they are easy to forget and easy to lose track of (Johnson, 2012). What is needed is a method for capturing these hunches, and a system that allows these hunches to be recaptured on demand. Johnson advocated keeping a “spark file”—a single document in which a person writes down all their ideas, be it for a journal article, an experiment, a design concept, or even a thought provoking quote. The key for Johnson was to not only keep the spark file, but also frequently go back and read through the entire document. In this way hunches from years ago that may finally be ready to mature can be brought back to a person’s attention. As Johnson stated, using a spark file is akin to brainstorming with your past selves.
Keller, Pasman, & Stappers (2006) conducted interviews with experienced designers in order to investigate the design process, and found that designers used a method similar to Johnson’s spark file. The designers relied on a physical collection of design concepts—stacks upon stacks of pictures. While sorting through these stacks of pictures looking for a specific one, designers often came across other pictures that spawned new and useful ideas. Interestingly enough, the designers also had digital collections of images. However, when the image was on a computer, designers became engrossed in finding the exact right picture on the computer, eliminating the fortuitousness of sorting through physical images. To address this problem, Keller, Visser, van der Lugt, and Stappers (2009) developed a tool called the Cabinet (pronounced Cab-in-eigh)—a workbench that allowed designers to easily manipulate, access, and store digital design images. The Cabinet supported serendipitous encounters by displaying random images while idling, and provided a digital method for quickly rifling through images.

Just as Magnus Carlsen’s chess ability and Kobe Bryant’s basketball skill are often credited to innate talent, good designs seem to simply “appear out of the blue”. This neglects the role that practice, the desire to practice, and the type of practice play in the design process. Unfortunately, just as high schools and college focus solely on teaching essay writing instead of creative writing, human factors programs tend to focus solely on teaching scientific theory at the expense of teaching design. Future human factors professionals should graduate with not only a solid understanding of the scientific
method, but also an understanding of good design practices (such as keeping a spark file or stacks of interesting design concepts).

The following section describes the development of HAI interfaces that enable a single operator to use multiple UVs to support base defense missions. This section focuses on the EID principles that were followed throughout the design process as well as empirical evaluations of initial HAI concepts. However, the role that art and serendipity played in the design process is also discussed.

*Spanning the Gulf of Execution. Rule-Based Interaction. Direct Manipulation.*

“In defense, it is not so much machines that should be put at the ready, but strategies” - Vitruvius 10.16.8

In Chapter I it was suggested that in order to bridge the gulf of execution, HAI must have controls and displays consistent with a human operator’s reasoning capabilities. Rasmussen’s SRK framework was proposed as a method for understanding alternative ways to represent constraints, signals, signs, and symbols, which in turn, distinguish three levels of human performance, skill-based, rule-based, and knowledge-based. In the context of supervisory control, an example of skill-based behavior is a UV operator manually controlling a sensor to keep a moving target in view. The location of the target in time and space provides the operator with the signal, and his or her goal is to

---

13 From Vitruvius’ *De architectura* (On architecture) as cited in Rowland, Ingrid (2014). Original Latin: *ita in repugnatoriis rebus non tantum machinae sed etiam maxime consilia sunt comparanda.*
continuously adjust the sensor any time the target deviates from the center of the screen. Additional skill-based behaviors associated with supervisory control include manually flying, driving, or steering a UV and maintaining a specific heading, altitude, or speed. Skill-based behavior is feasible only if there are minimal time delays; as the time lag between an operator command and resultant action increases, skill-based control becomes impractical.

In rule-based behavior an operator has a set of predetermined solutions that are triggered by specific conditions (i.e., signs). Figure 3 illustrates rule-based behavior for determining the best search pattern to use in the presence of specific combination of signs for a UV operator to find a target. For example, if an operator is searching in a non-mountainous region for a target that is likely to still be near the last known location but whose direction of travel is unknown, a sector search is the best option. These rule-based solutions often help an operator quickly identify a “good enough” solution.

Knowledge-based behaviors occur when an operator encounters a novel unexpected situation for which no procedure exists (Rasmussen, 1983). In these instances an operator may evaluate and critique skill and rule-based tactics, in order to learn from mistakes and to take advantage of opportunities and avoid threats that arise due to changing situation contingencies. For example, an operator may modify a rule-based strategy to find a target based on his or her knowledge of insurgent activity in the area or the limitations of the only UV he or she has available.
In order to successfully bridge the gulfs of execution and evaluation the user interface must support all three modes of interaction (skill, rule, knowledge-based). One approach to designing such an interface is EID. EID emphasizes developing interfaces and automated systems that promote coordination between the human operator and the automatic system by representing constraints (Borst, Flach, and Ellerbroek, 2015). Since its inception over 25 years ago, the majority of EID interfaces have focused on providing support for manual control tasks, which typically involve skill-based behaviors (Borst, Flach, and Ellerbroek, 2015). The goal here is to design interfaces and automated tools that effectively support rule-based and knowledge-based behaviors as well.

In a study focused on developing a cognitive model of air-to-air combat pilots, Amalberti and Deblon (1992) found that preparation was crucial to mission success. During pre-mission planning pilots would examine each mission leg, identify potential threats, and develop strategies to address these threats. As a result, many problems that a pilot might encounter have been solved before the pilot even takes off. Similarly, base defense personnel have developed QRCs that list the actions to take in response to potential security threats such as fence alarms, mortar attacks, vehicle-borne improvised explosive devices, etc. These QRCs are rule-based behaviors that enable security forces to quickly respond to rapidly changing events and muddle through with the aid of skill- and knowledge-based adjustments.

During the CTA interviews described in Chapter II, the security force personnel described the role of the BDOC commander as similar to the role of a head coach on the
Just as a head coach is responsible for calling and modifying plays in real time, delegating authority to his players at times (i.e., allowing the quarterback to audible in certain situations), and ensuring the correct personnel is used (i.e., putting in the strong pass rusher on passing downs), the BDOC commander is responsible for modifying and tailoring mission plans in real time and ensuring plans are updated in response to dynamic situations and reassigning assets as needed. Using this analogy, Miller & Parasuraman (2007) designed a HAI called the Playbook® (see Figure 6), which allows an operator to call a play from a predefined play library at various levels of abstraction to issue commands to the autonomous players (Miller & Parasuraman, 2007).
While the Playbook® concept is a great starting point, the majority of Playbook® interfaces reviewed were text-based and menu driven and did not support direct manipulation—controls and displays that are consistent with an operator’s reasoning capabilities (Bennett & Flach, 2011). Though various solutions were considered an unlikely source proved the most fruitful—video games.

City of Heroes is a now defunct massive multi-player online role playing game (MMORPG) which placed gamers in the role of a super hero or super villain. One character class gamers could choose was the Mastermind, whose main super power was
the ability to direct up to six henchmen/pets—of varying abilities—to achieve their goals.
The Mastermind controlled these henchmen/pets via the interface shown in Figure 7.

Figure 7. City of Heroes Mastermind Interface (NCSoft, 2012. Author’s Screenshot).

Note the set of three icons that repeat. The first icon (the crosshair) instructs the
henchmen/pet to attack an enemy. If the Mastermind clicks this button, whatever enemy
the Mastermind has targeted is attacked by the henchmen/pets. Clicking the second
button (the arrow) and then clicking a location in the virtual world sends the
henchmen/pets scurrying to that location. The third button (the human figure) commands
the henchmen/pets to attack the enemy that is dealing the most damage to the
Mastermind. Also, note that this interface allows the Mastermind to command a specific
henchmen/pet, all henchmen/pets of a certain type (the Dire Wolf, all Lions, or all
Wolves), or all henchmen/pets. With this interface, the Mastermind can focus on
accomplishing tasks—eliminating bad guys, moving to specific locations, eliminating the biggest threat—rather than on controlling each henchmen/pet individually.

The City of Heroes interface allows the Mastermind to communicate high level goals—move to a location, focus force on a specific enemy, eliminate the biggest threat—while providing the henchmen/pets the flexibility to accomplish these goals in the manner they see fit. To put it into military terms, the interface allows the Mastermind to communicate command intent—a concise statement that dictates a commander’s desired end state to his or her subordinates (Shattuck, 2000). In practice, the conciseness and detail of commander’s intent varies (see Table 3).

Table 3. Commander’s Intent. (Shattuck 2000).

<table>
<thead>
<tr>
<th>Detailed Commander’s Intent</th>
<th>Highly flexible Commander’s Intent</th>
</tr>
</thead>
<tbody>
<tr>
<td>The purpose of X Brigade’s operation is to protect the Corps, rear and build-up of follow-on friendly forces. In support of Division and Corps, we must attack rapidly to the west in the Central Corridor, destroy the lead motorized rifle battalion (MRB) of the XXX Motorized Rifle Regiment (MRR) between Phase Line (PL) IMPERIAL and PL EXCALIBUR, and then seize defensible terrain along PL EXCALIBUR. To do this, X-X Infantry (Light) will infiltrate to secure Hill 780 (NK4411), deny the enemy its use, and block to the west to prevent the enemy’s use of the mobility corridor between Hill 780 and the south wall of the Central Corridor (Avenue of Approach 3). Task Force X-XX, the brigade main effort, will move to contact in zone, fix the advance guard main body (AGMB) and destroy it with an enveloping attack in depth. Brigade deep artillery fires, close air support and scatterable mines will be designed to attrit its commitment into the Brigade zone, and force the AGMB into the southern avenue of approach,</td>
<td>OK, you want your brigade commander’s priority? Take care of this. If you don’t get this right then TF X-XX will not be able to get through.</td>
</tr>
</tbody>
</table>
Shattuck not only suggested that a concise statement that provides subordinates with flexibility is key for mission success, but also indicated that commanders need to be able to specify detailed courses of actions to ensure synchronicity among subordinate units. An interface that allows a commander to communicate intent at various levels of abstraction as required is needed.

IMPACT’s Play Calling interface has been designed to reflect this approach. In order to respond to a suspicious vehicle the operator doesn’t have to assign a specific UV. Instead, the operator can simply click the location of the suspicious vehicle and hit the “Air Inspect Point/Object X” play icon. That’s all the IA needs to recommend a course of action.

The icons used in IMPACT Play Calling interface were inspired by City of Heroes. Within the City of Heroes interface, icons are used to pictorially communicate specific capabilities, and classes of capabilities are color coded. Figure 8 shows the interface for a villain whose primary superpowers involve the ability to control plants (green icons) and to attack enemies using plant thorns (pink icons). The icons also contain information about the particular power each icon represents. For example, in the
bottom row (Row 1), Icons 5, 7, and 8 all have an arrow pointing to the right—indicating a ranged attack—while Icons 1 and 2 have an encased humanoid figure indicating a hold—a power that prevents an enemy from taking any action. Additionally, the player has arranged the icons meaningfully within the rows. Row 1 contains powers that attack enemies while Row 2 mainly contains powers that impact the villain—selecting Icon 1 turns on super speed, Icon 6 teleports the villain, and Icon 7 increases the villain’s accuracy. In the top row—the row labeled 4—the player has placed powers related to summoning henchmen/pets in the upper right corner.

Figure 8. Mastermind Interface. (NCSOFT, 2012. Author’s Screenshot).

This approach inspired the development of IMPACT’s Play Calling interface (see Figure 9). Each icon provides information about the type of unmanned vehicle assigned to the play as well as the play type. Vehicle type is indicated by the presence and location of a shape—a diamond in the upper left corner for UAVs, a rectangle in the lower left for UGVs, and a pentagon in the lower right for USVs. Play type is indicated by the symbol in the icon’s interior—a plus sign for a point search, a line for a route search, a square for an area search. Additionally, plays are assigned to the rows by type—point search plays in the top row, route search plays in the second row, area search plays in the third row,
friendly vehicle plays in the fourth row, and hostile vehicle plays in the bottom row.

Mersch, Behymer, Ruff, & Calhoun (2016) found that color coding these icons did not improve operator accuracy or response time in terms of finding the correct play. However, they did find that arranging the icons by vehicle type — all UAV plays in Row 1, UGV plays in Row 2, USV plays in Row 3, etc. — significantly improved response time over arranging the icons randomly or by play type. Nevertheless, because this result may have been due to the wording of the trial prompts used — prompts always specified the UV type first — an organization by play type was retained for the validation phase.

Figure 9. IMPACT Play Calling Interface.

The IMPACT Play Calling interface not only allows the operator to communicate intent at a higher level of detail if desired (e.g., assign specific vehicles, altitudes, airspeeds, standoff angles, etc.), but also allows the user to quickly assign high level tasks for the autonomy to execute helping to bridge the Gulf of Execution. The Play Calling interface lies directly on a base map that displays the UVs as well as locations/targets. It
supports direct manipulation by allowing operators to call plays directly from the map by interacting with a UV or with a location/target. The Play Calling interface also enables an operator see the play options that make sense for a specific vehicle or location/target. For example, in Figure 10, right clicking Point Alpha — a sea-based point — provides point plays for UAVs and USVs, but eliminates route plays, area plays, and UGV plays. Similarly, right clicking on a UGV provides a list of UGV plays, eliminating UAV and USV only plays.

In addition to the Play Calling interface, operators can use an interface component called the Task Manager to call plays (see Gutzwiller & Lange, 2016 for an in-depth description of the Task Manager). In order to support rule-based behavior, the Task Manager constantly monitors the chat window for incoming tasks that the autonomous system can link to an existing QRC. For example, in Figure 11, the Task Manager — in response to a chat message stating that an unidentified boat sighted at Point Sierra is
headed towards the coast — has recommend that the operator get eyes-on the watercraft using a point search play and surround the watercraft using a blockade play.

Figure 11. Task Manager.

The Play Calling interface and Task Manager both enable the operator to quickly and efficiently take appropriate action in response to mission events. When a play is called the operator’s intent is communicated to the autonomy and the autonomy generates a set of plans to accomplish the operator’s goals. With this complete, the next goal is to build the bridge to cross the Gulf of Evaluation: how can the operator assess and evaluate the autonomy’s plan(s)?

**Spanning the Gulf of Evaluation. Direct Perception.**

An oft-cited goal of autonomous system developers is transparency, which Miller (2014) claimed has been interpreted as an autonomous system should be glasslike in that
its functions, behaviors, and rationale are “see through” to the operator. Miller made a
distinction between this view of transparency and practical transparency, which he
defined as conveying to the operator the information that affords comprehension. For
IMPACT, presenting the process algebra (see Figure 12) that drive the CCA’s route
planning or the CDOs that enable the IA to recommend UVs to fulfill tasks would
obviously not provide practical transparency.

Figure 12. Process Algebra (Varricchio, Chaudhari, & Frazzoli, 2014).

In response to an operator play call, IMPACT’s IA recommends a plan after
considering the plan’s probability of success (sensor, speed, detectability), the time in
which it can be accomplished, the amount of resources the plan consumes, and the impact
the plan will have on other ongoing tasks. To achieve practical transparency the user
interface should communicate the IA’s recommendation in a manner that, in Miller’s
terms, affords operator comprehension. A review of the literature provided several potential solutions. One of the most promising was Findler’s (2011) Visual Thinking Sprocket (see Figure 13).

![Figure 13. Visual Thinking Sprocket (Findler, 2011).](image)

This design was adapted (see Figure 14) to communicate a given plan’s quality to the operator (Behymer, Ruff, Mersch, & Calhoun, 2015). This sprocket conveys several pieces of information to the operator. The size of each pie piece represents the weighting that each parameter was given by the agent. For example, estimated time enroute (ETE) was given the most weighting, followed by probability of detection (PoD). The parameters were ordered based on their priority in a clockwise fashion, with the highest priority parameter (ETE) located at the twelve o’clock position. Color is used to represent
a plan’s status. If everything is going according to plan (e.g., the UV is still expected to arrive at task location on time) the middle circular segment of that parameter’s pie slice is green and indicates a “normal/ideal” state (see Impact parameter in Figure 14). A warning state is represented with three yellow segments and indicates a “slight” deviation from the ideal state. An error state is represented with five red segments and indicates a “severe” deviation from the ideal state. The specific location of the segment with the brighter, more saturated color indicates whether the value exceeds or is less than the desired operating range for that parameter. For example, in Figure 14 the UV is now expected to arrive slightly ahead of the scheduled ETE time (bright yellow segment near pie’s center) and with a greatly reduced probability of detection (PoD) than expected (bright red segment near pie’s center). This could notify the operator that the UV could possibly get closer to the target for increased sensor quality (since PoD is critically lower than nominal) or could possibly be re-planned to image another target while enroute.

Figure 14. Representing Play Quality (Behymer, Ruff, Mersch, & Calhoun, 2015).

Based on the findings of Cleveland and McGill’s (1985) graphical perception task study in which participants were much better at detecting differences in length (e.g., the
width of a bar) than differences in angle or area (e.g., the angle/size of the pie wedge), an alternative bar chart visualization was created (see Figure 15). In the bar chart, the parameters were ordered from left to right based on their priority, with the highest priority parameter at the far left. The width of each column also represented the priority of each parameter. The coding of the color of each segment of the bars was similar to that employed in each segment of the slices in the pie chart. Behymer, Ruff, Mersch, & Calhoun (2015) found that participants both preferred and were faster at retrieving parameter state and priority information from the bar chart as compared to the sprocket design.

Figure 15. Play Quality Bar Chart (Behymer, Ruff, Mersch, & Calhoun, 2015).

While the bar chart provides a way for an operator to quickly grasp the quality of a given plan, it is not optimal for comparing across plans. When the operator calls a play, IMPACT’s IA generates Pareto solutions. A Pareto solution is a set of efficient solutions—rather than an optimal solution—in response to a multidimensional problem (Legriel, J., Le Guernic, C., Cotton, S., & Maler, O., 2010). For example, one solution may optimize time while another solution may optimize fuel. One method that has been
suggested for visualizing the Pareto solution space is the Pareto Front. In Figure 16 many plans (each represented by a gray box) are mapped on a three axis graph (minimizing time to arrive, maximizing information, and minimizing bandwidth usage). While the Pareto front certainly allows an operator to compare multiple plans, it is limited to three dimensions. What the operator truly needs is an efficient way to represent N plans across N parameters.

Figure 16. A Pareto Front.

While a literature search focused on designs to compare plans proved fruitless, the author came across the graphic shown in Figure 17 in an article focused on USV capabilities (Savitz, et al., 2013).
Figure 17. Comparing USVs across Many Parameters (Savitz, et al., 2013).

While having nowhere near the societal impact as Archimedes’ seminal moment (or even Douglas Caldwell’s), the author certainly experienced a feeling of Eureka on seeing Figure 17. One can quickly determine the best USV for a specific parameter (The ACTUV can carry the biggest payload) while also seeing how each USV compares across all relevant parameters. This display—called a parallel coordinates plot—was adapted to visualize the IMPACT AI’s Pareto solution space. The Course of Action (COA) Planner, shown in Figure 18, also allows an operator to quickly determine the best plan for a specific parameter.
Each parameter is assigned a column. Parameters are ordered by importance, with the leftmost being the most important. The overall quality of each plan is the leftmost column and is distinguished from the individual parameters (that are combined to determine the overall quality based on weightings assigned to each parameter). Each plan is assigned a color. The quality of each plan for that specific parameter is mapped onto the parameter’s column. For example in Figure 18, Plan #1074 is the best for ETE and Plan #1073 is the best for Force (capability to act as a deterrent), CC (crowd control capabilities), and Track (tracking capabilities).

Behymer, Mersch, Ruff, Calhoun, and Spriggs (2015) conducted a study that compared the COA Planner to a series of bar charts like the one shown in Figure 15—if the COA Planner showed three plans, three bar charts were shown, one bar chart for each...
plan. Participants both preferred and were faster at comparing across plans when using the COA Planner as compared to the series of bar charts.

An additional concern facing the design team was the possibility of the autonomy returning a null solution set (i.e., the system cannot recommend a play based on the operator’s constraint selection). The system should communicate to the operator why it cannot develop a plan and provide alternatives. For example, if an operator calls a ground inspect at Point Alpha, but Point Alpha is beyond a UGV’s communication range. IMPACT’s IA, rather than saying “Ye cannot call a ground inspect at Point Alpha”, communicates the problem and recommends calling a multi-vehicle play where a UAV acts as a communication relay to extend the UGV’s range allowing it to travel to Point Alpha. Additionally, if the play the operator wants to call is a play that requires an asset that is currently on another task, IMPACT allows the operator to queue the play and set it to activate as soon as the asset becomes available. Finally, a “Hammer of Thor” feature allows the operator to ask the autonomy to free up any assets needed in the event of a high priority task by pausing any conflicting ongoing plays.

Once a play is called, it is added to the Active Play Manager (see Figure 19), which contains a list of all ongoing plays. The Active Play Manager lists the name and location of each play along with which UVs are on the play. Each play in The Active Play Manager has a column on the far right that is color coded to provide a high level overview of the play’s status. For both plays in Figure 19 the green column is filled in
indicating these plays are progressing according to plan. If the Rainbow autonomies framework detects a deviation from the plan—a headwind slowing the UAV down enough to delay its arrival time—the yellow or red columns will fill in depending on the delay’s severity. The operator can click on the play to open up the bar chart shown in Figure 15 to get more information about the problem.

Figure 19. Active Play Manager.

**Supporting Knowledge-Based Behavior.**

While the Play Calling Interface and the COA Planner represent a good start at bridging the gulfs of execution and evaluation, the real benefit comes from connecting the two concepts to allow the user to directly manipulate play calling variables and see how the IA’s recommendation changes in response. A great example of this type of functionality can be found in the work of McEwen, Flach, and Elder (2012), who
designed a user interface to allow patients and doctors to see how a variety of health factors—cholesterol, blood pressure, smoking habits—impact cardiac health. While the visualization of how these factors impact cardiac health is impressive, the true power of the display comes from allowing the patient or doctor to demonstrate how a change in one variable—like quitting smoking—can impact cardiac health. For example, in Figure 20, a 46 year old smoker has a medium level (13.4%, yellow text) of cardiac risk.

Figure 20. Medium Level of Cardiac Risk (McEwen, 2012).
By clicking the “No” button next to “Smoke” the Doctor can immediately convey to the patient that quitting smoking will halve their risk of cardiac disease (7.2%, colored green) (see Figure 21).

Figure 21. Low Level of Cardiac Risk (McEwen, 2012).

Similarly, the IMPACT interface allows the operator to (1) provide the IA with information that will help it develop a better plan; and (2) change these variables quickly to see how it impacts the IA’s recommendations. For example, in Figure 22, the operator can select the size of the target he or she is searching for (which impacts the altitude a UAV would search at and/or the zoom level of the sensor), environmental conditions that
impact the UVs (if it’s cloudy an IR sensor might be better than an electro-optical [EO] sensor), optimization parameters (ask the IA to optimize on time, fuel use, detectability, presence, crowd control, or tracking), or priority by selecting the corresponding icon. For example, in Figure 22 the operator has selected an environment condition of sunny (the sun icon is selected), has asked to IA to optimize on time (the stopwatch icon is selected), and has set priority to high (the icon labeled “HI” is selected). The operator can also specify that the UV conducting the play has a specific payload (e.g., synthetic aperture radar [SAR] sensor, non-lethal weapon) by selecting the corresponding sensor or weapon icon. With this interface the operator can quickly tell the autonomy to “Find the target while keeping a low profile (selecting the eye icon), optimize on time (selecting the stop watch icon), and it’s cloudy so consider that (selecting the cloud icon)”.

Figure 22. Play Workbook.
Additionally, a voice querying system was developed to support an operator’s ability to ask the autonomy “What If” questions. For example, using the speech recognition system, an operator can ask questions like “How soon can an IR sensor get to Gate 4?” or “How soon can a UGV get to Building 42?” The autonomous systems work in conjunction to provide an answer to the operator via the chat window as well as via speech over the operator’s headset.

Summary

While autonomous capabilities are necessary to support the DoD’s goal of having a single operator control multiple unmanned vehicles they are not sufficient. An interface that supports skill-based, rule-based, and knowledge-based behaviors is vital for mission success. To this end, multiple autonomous components have been developed. CCAs rapidly calculate optimal vehicle routes based on situational constraints. IAs recommend plans based on situational factors as well as in response to operator requests. The Rainbow autonematics framework monitors the ongoing situation and alerts the operator to errors. These autonomous tools along with their associated visualization methods provide the operator with the ability to smoothly move between skill, rule, and knowledge based control as required by the situation. The ensuing section showcases a brief vignette in which the IMPACT operator works with IMPACT’s autonomous systems, through the HAI, to complete a mission task (see Calhoun, Ruff, Behymer, & Frost, accepted for a complete description of the IMPACT HAI). A flowchart of this entire vignette can be found in Figure 29 at the end of the vignette.
**IMPACT Vignette**

In this scenario, the IMPACT operator has just received a chat message from the BDOC that an unidentified watercraft has been detected at Point Sierra (see Figure 23).

![Figure 23. Chat Message Alert of an Unidentified Watercraft.](image)

The Task Manager detects this chat message and realizes that it implies a task for the IMPACT operator. The Task Manager adds the task to the Task Manager window, and provides the correct courses of action (taken from the operator’s QRC for unidentified watercraft) (see Figure 24). In this case, step one in the QRC is to get eyes on the watercraft with a point search play (either using a UAV or USV as Point Sierra is in the water), step two is to continue surveilling the watercraft until the all clear is given and blockade the watercraft if applicable (i.e., if the watercraft is described as heading towards the shore, which in this case, it isn’t).
Figure 24. Task Manager Provides QRC for Unidentified Watercraft.

The operator then calls an Point Search at Point Sierra by either clicking the Point Search icon from the Play Calling Interface (fifth icon in the first row of Figure 9) or right clicking on Point Sierra and selecting the Point Inspect icon from the Radial Menu (see Figure 10). The IA responds to this play request by asking the CCA to calculate how long it will take each UAV and USV to reach Point Sierra (UGVs are not considered because Point Sierra is in the water). The CCA returns an estimated time enroute for each UAV and USV to the IA, and the IA selects the UAV or USV which can reach Point Sierra in the least amount of time (in this case FN-43). Note that even though SY-10 is physically closer to Point Sierra (see Figure 25), FN-43 is faster, and can reach Point Sierra sooner even though it is further away. The IA then asks the CCA to plan a route from FN-43’s current location to Point Sierra. This route plan is given a color (in this
case salmon) and is represented on the map with dashed lines for the operator to review (see Figure 25).

Figure 25. Plan for Responding to Unidentified Watercraft, FN-43 to Point Sierra.

Once the operator verifies that this plan is acceptable, he or she executes the plan by clicking the checkmark in the Play Workbook. The play is added to the Active Play Manager, FN-43’s icon and current route change to salmon, and the dashed route line becomes solid to reflect the plan is now executing (see Figure 26). Note that if this had been a multi-vehicle play like a blockade, all UVs on the play would have the same color.
Figure 26. FN-43 Flies to Point Sierra to Investigate the Unidentified Watercraft.

As FN-43 flies to Point Sierra, the Rainbow autonomies framework monitors its progress and alerts the operator to any unforeseen events (e.g., engine failures, sensor failures, restricted operating zones). In Figure 27, Rainbow has detected that FN-43’s sensor has failed and alerts the operator by placing a red circle around FN-43. The operator can get more detailed information about the error by rolling over FN-43’s icon with the mouse cursor.
To explore other options to investigate the unidentified watercraft, the operator clicks on the play in the Active Play Manager to edit the play. As the IA isn’t aware of FN-43’s sensor failure yet, the operator opens up the COA Planner to view other potential courses of action. The operator quickly sees that SY-10 has the next quickest ETE to Point Sierra, and calls this play directly from the COA Planner by clicking the play button (see Figure 28).

Figure 28. COA Planner Compares Plans to get to Point Sierra.
Next, the operator receives a chat message from the BDOC asking how long it will take to get eyes-on the unidentified watercraft at Point Sierra. The operator queries the system with a voice command: “How soon can FN-43 get to Point Sierra?” The IA detects this query and sends a request to the CCA to calculate the time it will take FN-43 to get to Point Sierra and returns this information to the operator via chat message as well as voice over the operator’s headset.
**Summary**

This chapter described IMPACT’s HAI and how it was designed to bridge the coordination gulf between the human operator and IMPACT’s autonomous systems, enabling effective collaborations in support of UV operations. The next step is to determine if these interfaces can improve performance in the context of a base defense mission. The next chapter focuses on determining an appropriate experimental methodology to evaluate IMPACT’s HAI.
CHAPTER IV. Validate

“We create models to explain nature, but the models wind up gatecrashing nature and driving away the original inhabitants”. (Mitchell, 1999)

Flach, Schwartz, Courtice, Behymer, and Shebilske (2010) argue that the first thing researchers must do is identify the system of interest. The goal of this research is to design HAIs that enable effective collaboration between human operators and autonomous tools. To achieve this goal, the experiment must support multiple performance measures in an experimental setting that, to the extent possible, captures and represents the constraints of the natural work domain. Flach et al. (2010) call this type of experimental setting a synthetic task environment, the key to which is not the physical fidelity of the simulator, but how well the research question maps to the properties of the real-world mission environment. Conducting experimental studies in the context of the natural environment has been labeled a macrocognitive approach (Flach et al., 2010), and as it is categorically different from traditional psychological approaches, further discussion is warranted.

In his classic psychology textbook William James (1892) lamented psychologists who believe that a system can be understood by breaking it down into its components and studying these components in isolation. James compared this approach (which he labeled brass instrument psychology) to trying to understand a constantly changing river by examining buckets of staid river water. Change is the defining characteristic of a river,
just as change was the defining characteristic of James, who had a permanent and insatiable craving for it (Richardson, 2006). As James (1909) put it, “The essence of life is its continuously changing character; but our concepts are all discontinuous and fixed”. James’ views are echoed by Egon Brunswik, who argued that psychological processes are adapted to the environments in which they function and stressed the importance of retaining the actual characteristics of the environment during experimentation (Dhami, Hertwig, & Hoffrage, 2004).

Unlike the traditional psychology approach that seeks to determine cause and effect by isolating components of behavior or mental life in the laboratory, Flach and Voorhorst (2016) discuss the importance of context using Simon’s (1969) parable of an ant walking on a beach. The ant’s path is a result of the complexity of the beach, not the complexity of the ant. Just as important, as the beach is changing the ant, the ant is changing the beach, or as James (1907) put it, “The novelty soaks in; it stains the ancient mass; but it is also tinged by what absorbs it.” The system is shaped by the components that compose it, but the components are also shaped by the system. By failing to account for this, traditional psychology methods are inadequate in addressing real world complexities.

The Tyranny of the Quantifiable

Context matters, and yet, psychologists frequently try to eliminate context in the name of experimental control. In fact, psychologists often bow to the tyranny of the
quantifiable, a term coined by Solnit (2014) to describe the way that what can be measured usually takes precedence over what cannot, especially when it comes to complex, subtle, and fluid phenomenon. Pirsig (1991) provided an excellent example of this kind of neglect, when he described walking through a Native American reservation with the tribe’s chief, a professor of anthropology, and a woman from the Association of American Indians. Pirsig spoke of how a dog had been following the group and the woman had asked, “What kind of dog is that?” and the tribe’s chief had replied, “That’s a good dog”. Pirsig elaborated:

“Laverne had been asking the question within an Aristotelian framework. She wanted to know what genetic, substantive pigeonhole of canine classification this object walking before them could be placed in. But John Wooden Leg never understood the question. That’s what made it so funny. He wasn’t joking when he said, “That’s a good dog”. ... The whole idea of a dog as a member of a hierarchical structure of intellectual categories knows generically as “objects” was outside his traditional culture viewpoint. What was significant, Phaedrus realized, was that John had distinguished the dog according to its Quality, rather than according to its substance.” (Pirsig, 1991)

It’s easy to weigh the dog or measure the dog’s ears or note the color and thickness of a dog’s fur and determine its breed. It’s harder to determine if the dog is in fact a “good dog”. The end results of the tyranny of the quantifiable is often experimental
findings that have little to no pragmatic value to aiding our understanding of real world phenomenon.

For example, laboratory experiments can be designed to showcase how poor humans are at decision making by measuring their performance at logic puzzles. Flach and Voorhorst (2016) discussed the Wason Task (Figure 30), in which participants are asked to determine which cards need to be turned over to determine the validity of the statement, “If a card has a consonant on the letter side, then it has an odd number on the number side”. The majority of participants turned over the D (correct) and the 3 (incorrect). It doesn’t matter if there is a consonant or a vowel on the back of the 3 card, because either result doesn’t invalidate the rule. The other card that must be examined is the 2 card, because if the 2 card has a consonant on the back, then the rule is violated.

![Figure 30. Wason Task (1966).](image)

Yet, if participants are given the same logical task, but with context, performance improves dramatically (Cosmides & Tooby, 1992). In Figure 31 the task is to determine which cards to flip over to validate the rule, “If a person is drinking beer, then the person is over 21”. In this example, participants perform well, identifying that they need to know the age of the person drinking the beer as well as what the 18 year old is drinking.
The key here isn’t that participants succeed because they have prior experience identifying underage drinkers. The key is that people have a schema for permission and detecting cheaters. For example, participants also do well in determining which cards to flip over to determine the validity of the rule, “If a person is eating cassava root, then they must have a tattoo” (Figure 32). Despite having no previous experience with this particular rule, participants know they need to determine if the person eating cassava root has a tattoo and if the person without a tattoo is eating cassava root.

At this point, the reader may be wondering, what does an ant on the beach or card games have to do with UV operations? **Everything!** \(^{15}\)

---

\(^{15}\) This particular phrasing pays homage to Henry James, the great American novelist and younger brother of William James. James used this technique (a question followed by an emphatic “Everything!”) in many of his novels, perhaps most famously in *The Portrait of a Lady* when Isabel Archer asks Madame Merle, “Who are you what are you? … What have you to do with my husband?.... What have you to do with me?” and Madame Merle replies, “Everything!” With this utterance Isabel learns that her marriage, which she thought she entered into freely and serendipitously, had been manipulated into being by Madame Merle.
"Laws may help you hack through the jungle, but no law changes the fact you’re in a jungle" (Mitchell, 1999)

In April 1994 in the Northern Iraqi no-fly zone, two U.S. Army Black Hawk helicopters were accidently shot down by two U.S. Air Force F-15 fighter jets, killing 26 military and civilian personnel (Snook, 2000). The proximal cause of this friendly fire incident was determined to be the misidentification of the Black Hawks as Iraqi Hinds by the F-15 pilots (Snook, 2000). However, this is just the beginning of the story.

Figure 33 lists all the factors Snook (2000) identified as important in understanding how this incident could have occurred. This figure emphasizes the complexity of the context that the incident occurred in. Yes, the F-15 pilots misidentified the Black Hawks as Hinds, but the F-15s were also not expecting any friendly forces to be in the no-fly zone because they hadn’t received any information from AWACS, which happened to have an undermanned crew that day, which was the result of a smaller defense budget, which occurred because of the collapse of the Soviet Union.
Figure 33. Factors Leading to the Shootdown. (Snook, 2000).
When placed in this order it may appear that a linear sequence of events led to the disaster, but as Flach and Voorhorst (2016) discuss, it’s really a set of nested constraints, with each level providing the context to understand the constraints at the level beneath it (see Figure 34).

![Abstraction Space](image)

**Figure 34. Abstraction Space (Flach and Voorhorst, 2016).**

Note how Snook (Figure 35) uses a similar method to map out the levels of abstraction for the friendly fire incident. Using this method, researchers can gain a deeper understanding of the context that is critical for identifying methods for preventing future friendly fire incidents. The researcher can also use this information to design test environments in which to evaluate any assessment, training, or tools designed to improve operator performance in these situations. Unfortunately, traditional psychology experimental approaches often seek to demonstrate the errors humans make compared to mathematical logic in toy world problems devoid of the rich context people use in everyday decision making.
Evaluating IMPACT’s HAI

In order to support a macrocognitive approach, each experimental trial in this research utilized a synthetic task environment in which a single operator used multiple UVs to support base defense operations over the course of an hour-long mission. The mission scenario and operator tasks (described in-depth in Chapter V) were derived from the CTAs with security force personnel described in Chapter II, and the performance measures (described in-depth in Chapter V) were derived from what security force personnel deemed important for mission success. The emphasis wasn’t on experimental control or designing an experiment that would yield clean results for easy statistical analysis. Rather, the experiment was designed to evaluate IMPACT’s HAI in the context of a simulated base defense mission that had as many real-world elements as possible.
Additionally, one of the biggest mistakes autonomous system developers make when evaluating their tools is to compare the performance of the autonomous system against the performance of human operators. Recall the DARPA coordinators project discussed in Chapter I in which a team of human operators significantly outperformed intelligent agents. A better evaluation is to compare the performance of human operators to the performance of human operators WITH the automated system. In order to demonstrate the effectiveness of IMPACT’s HAI this research compared IMPACT to a Baseline condition that represented the current state-of-the-art at the beginning of the IMPACT project. The Baseline condition had a subset of IMPACT’s capabilities including the CCA to assist in route planning and a HAI to interact with the CCA. However, the Baseline condition lacked IAs to support vehicle recommendations, the Rainbow autonamics framework for plan monitoring, the task manager, and the voice recognition system. Operator performance and overall mission effectiveness was expected to be significantly better with IMPACT as compared to Baseline. Additionally, participants were expected to prefer IMPACT over Baseline.

Finally, this research also was designed to investigate the extent to which IMPACT’s HAI can aid performance as the complexity of the mission increases. During the interviews with BDOC personnel it became clear that base defense operations are characterized by long periods of inactivity with sudden bursts of intense workload in response to threats. To this end, two levels of complexity were examined in the experiment, a low complexity mission and a high complexity mission (described in detail
in Chapter V). Operator performance was expected to be worse in the high complexity missions as compared to low complexity, but the performance decrement was expected to be less with IMPACT’s HAI than with Baseline.
CHAPTER V. Experimental Study Methods

Participants

Eight volunteers with relevant military experience participated in this study, four active duty and four who had previously served. Six participants had prior experience piloting UAVs (Global Hawk, Predator, Reaper, Scan Eagle, and Raven), one participant was a former Predator/Reaper SO, and one participant was an experienced security force and base defense expert. Seven participants were male (one female) and all participants reported normal or corrected-to-normal vision, normal color vision, and normal hearing. The average age of participants was 43.6 years ($SD = 10.84$).

Design

Tool. A 2 X 2 within-participants design was used, with each participant experiencing both Baseline and IMPACT at two different levels of task complexity. The order of conditions was counterbalanced by tool and task complexity. In the Baseline condition participants had access to the CCA and a HAI to work with the CCA. The IMPACT condition had these features as well as an IA to support plan recommendations, plan monitoring, task manager, voice commands, and associated HAIAs (Table 4).
Table 4. IMPACT and Baseline Differences.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Human Operator</th>
<th>HAI</th>
<th>CCA</th>
<th>IAs</th>
<th>Monitoring</th>
<th>Task Manager</th>
<th>Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>X</td>
<td></td>
<td>Subset of IMPACT capability</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IMPACT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

The key interface component of the Baseline condition was the Unmanned Systems Autonomy Services (UxAS) Wizard (see Figure 36), which allowed participants to interact with the CCA. The icons used in the UxAS Wizard used a similar encoding scheme as the Play Calling interface (with point plays having a plus sign within the icon, etc.). However, once a participant selected an icon from the UxAS Wizard he or she then needed to select a location as well as the specific UV(s) to assign to the task. For example, if a participant needed to perform a blockade of an unidentified watercraft at Point Sierra he or she first selected the blockade icon (the fourth icon in the fourth row in Figure 36), then selected Point Sierra, and finally selected all 4 USVs. The CCA would then plan routes for each of the USVs to set up a blockade of Point Sierra, which the participant could examine and then confirm by clicking a check mark. As with IMPACT, routes were represented by dashed lines during planning and solid lines when executing.
Complexity. Task complexity was varied by a combination of increasing the number and complexity of Random Anti-Terror Measures (RAMs) the participant needed to complete during the shift, increasing the number of commander queries the participant needed to respond to, increasing the amount of noise radio and chat chatter (i.e., messages that didn’t require participant action), and increasing the number of events (normal base defense, intruder, environment, UV faults) the participant encountered (Table 5). RAMs, commander queries, normal base defense events, intruder events, environmental events, and UV faults are described in-depth below in the scenario section. Chat chatter refers to chat messages from manned security patrols that didn’t require participant action (e.g., “Patrol 4: Beginning sweep of sector fifteen”). Radio chatter refers to pre-recorded radio calls from manned security patrols that didn’t require participant action (e.g., “Patrol Two is conducting a show of force at Gate 3”).

Figure 36. UxAS Wizard.
**Table 5. Task Complexity.**

<table>
<thead>
<tr>
<th>Scenario Difficulty</th>
<th>RAMs</th>
<th>Queries</th>
<th>Chat Chatter</th>
<th>Radio Chatter</th>
<th>Normal Base Defense Events</th>
<th>Intruder Events</th>
<th>Environment Events</th>
<th>UV Faults</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>4</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>6*</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td><strong>4</strong></td>
<td>3</td>
<td>4***</td>
<td>4***</td>
</tr>
</tbody>
</table>

*one required a communications relay
**one required a cordon or blockade
***one occurred during Intruder Event

**Apparatus**

The study was conducted in the Air Force Research Laboratory Supervisory Control and Cognition (AFRL/RHCI) branch’s state of the art Human Autonomy Laboratory (HAL) using the Fusion test bed (see Rowe, Spriggs, and Hooper, 2015 for an in-depth description of the Fusion test bed). The experimental configuration used in this study consisted of four stations, the C2 Operator Station, the Sensor Operator Station, the Test Operator Console, and the Simulation Station (see Figure 37). The Simulation Station used a Dell Precision T5400 and ran One Semi-automated Forces (OneSAF), a simulation tool that generated all friendly, neutral, unknown, and hostile forces during the experiment, with the exception of the UVs. The C2 Operator Station and Test Operator Console each used a Dell Precision T7910 while the Sensor Operator Station used a Dell Precision T5600. The C2 Operator Station, Sensor Operator Station, and Test Operator Console had identical monitor setups, with one Sharp PN-K322B 4K Ultra-HD LCD Touchscreen (3840 x 2160 resolution) and three Acer T272HUL LED Touchscreen (2560
Three Dell Precision R7610 located in a different room provided the sensor feeds for the UVs (four feeds per machine). SubrScene, an in-house simulation visualization toolkit was used to provide the sensor feeds.

Figure 37. Experimental Configuration.

Figure 38 shows the difference in screen layout between IMPACT and Baseline. In both IMPACT and Baseline, the top screen was a Tactical Situation Display (TSD) that provided a geo-referenced map with UV locations as well as UV-specific information (e.g., a UV’s current play, error indicators, a UV’s planned route, etc.). The TSD also had a Vehicle Summary Panel in both conditions. The Vehicle Summary Panel contained icons for each UV and allowed participants to quickly access additional information about a UV. In IMPACT the TSD also contained a window—labeled Voice.
in Figure 38—in which the participant received feedback from the voice recognition system. For example, if a participant asked a query via voice command, the answer appeared in this chat window. The right screen contained vehicle dashboards for each of the twelve UVs that provided vehicle state information (e.g., speed, altitude/elevation, payloads, sensor status) as well as the sensor feed. The bottom screen was called The Sandbox because in addition to displaying the same information available in the TSD, it allowed participants to generate plans for the UVs. In the IMPACT condition participants had access to the Play Calling interface, the Task Manger, and the Active Play Manager on this screen. In the Baseline condition participants had access to the UxAS Wizard. In both conditions the Sandbox contained the mission chat window, in which tasking from the simulated mission commander appeared. Finally, the left screen contained the Help window, which provided the participants with access to the QRCs, UV specification sheets, and voice commands. The left screen also contained the media manager, which allowed participants to view images taken by the SO.
Figure 38. C2 Station Layout for both IMPACT and Baseline Conditions.

**Scenario**

During the mission participants were placed in the role of an operator managing twelve UVs (four UAVs, four UGVs, and four USVs) to support base defense operations. The participant’s job was to use the UVs to accomplish tasks in response to requests by his or her commander which were generated from the TOC via pre-scripted chat messages. The participant had access to the UV sensor feeds but it was not his or her responsibility to monitor them; that role was performed by the Sensor Operator, who was played by one of the members of the experimental team. The participant’s main task was directing and monitoring the UVs in response to various events. For each event participants had a QRC available in the Help file that listed the correct response for that event. Events that could occur included:

1. **Patrols.** Participants were instructed that UVs currently not performing other tasks should patrol the base. Two patrol states were trained, normal full
coverage (used when there was no existing threat to the base) and highly mobile (used during intruder events).

2. RAMs. At the beginning of each mission participants were given a list of RAMs to complete. The RAMs list used for each scenario are provided in APPENDIX B. Participants were instructed to abandon RAMs during Intruder Events.

3. Normal Base Defense Events. During the mission participants were tasked to support various base defense operations. Participants were instructed to continue Normal Base Defense Events until instructed otherwise, even during Intruder Events.

4. Intruder Events: During the mission participants were tasked with responding to intruder events. Participants were instructed to continue responding to intruder events until the all clear was given.

In addition to these events, participants also had to respond to queries from their Commander via chat. Example queries include: What’s FN-42’s Altitude? How long would it take to get a Show of Force at Gate 3 in place? How many RAMs have been completed? Participants also had to respond to vehicle failures (e.g., sensor malfunctions, engine failures) and environmental events (e.g., restricted operating zones, dense smoke).
Four experimental scenarios were used, two low complexity scenarios and two high complexity scenarios. Each type (low and high) were matched, so that if an event occurred in the first low complexity scenario (Scenario A), an equivalent event happened in the second low complexity scenario (Scenario B) at the same time. Each experimental scenario was 60 minutes long and had an initial period of normal base defense operations lasting about 30 minutes, followed by an intruder event that lasted about 15 minutes, followed by a resumption of normal base defense operations for the final 15 minutes.

During the mission, the SO (played by a member of the research team) acknowledged participant actions and took images from the sensor feeds as required. For example, if the participant called a point inspect play to investigate an unidentified watercraft, the SO would take an image of the watercraft once the UV arrived and send it to the participant. The SO, depending on what the script called for (see APPENDIX C for a sample SO script), either gave the all clear after the image was taken (thus implicitly instructing the participant to return the asset to patrol) or stated that the all clear had not yet be given (thus implicitly instructing the participant to keep the UV on task). If the participant asked the SO about the status of non-SO related task, the SO would advise the participant to check his or her chat. For example, if the participant asked the SO if the unidentified watercraft had been imaged, the SO would reply with a yes or no. If the participant asked if the gate runner had been apprehended the SO instructed the participant to check his or her chat window.
The TOC operator, played by another member of the research team did not interact with the participant during the mission. However, the TOC was responsible for ensuring pre-scripted events occurred and injecting pre-scripted events as needed. For example, if the script called for the UV that was conducting a point inspect to lose its sensor feed, it was impossible to know which UV the operator would assign to the task a priori. Once the operator had selected the UV for the task, the TOC would disable that UV’s sensor feed on the fly (see APPENDIX D for a sample TOC operator script).

Procedure

The study took place over two days. On day one participants were trained on the base defense mission as well as how to use Baseline and IMPACT. Participants completed experimental trials on day two.

Day 1: Training. Participants were briefed on the goals and purpose of the study, signed an informed consent form, and were given a safety briefing. Next, participants completed a background questionnaire (see APPENDIX E) that collected basic demographic information (age, gender), unmanned vehicle operations experience, manned flight experience, and base defense experience. Participants were then seated at the C2 Operator Station to begin training. Training consisted of the lead experimenter instructing the participant how to perform specific actions using IMPACT and Baseline. Once a particular capability or function had been trained, the participant was sent a series of questions/tasks to accomplish via chat to ensure that he or she had been sufficiently
trained before moving on to the next topic. For example, once a participant had been trained on how to manipulate the map, chat messages were sent asking him or her to pan, zoom, and rotate the map using the mouse and the touchscreen.

Participants were first trained on the base defense mission they would be supporting, beginning with a description of the base and their role in the upcoming mission. Participants were also given a paper map of the base (see APPENDIX F) that they had access to for the duration of the experiment. The experimenter then provided a briefing on the capabilities of the UVs and instructed participants how to compare across the UVs to determine which UV to use in specific environmental conditions, when a specific optimization (e.g., optimize for crowd control) was required, or when a specific payload was needed. For example, FN-40 and FN-41 are faster and use less fuel, but FN-42 and FN-43 are less likely to be detected. This information was accessible throughout the experiment via the help menu, and the participant was also given a paper copy of this information (see APPENDIX G).

Next, participants were trained on the types of tasks they were responsible for performing (e.g., patrols, RAMs, normal base defense events, intruder event, commander queries, vehicle failures, environmental events) during the mission and the correct response for each. Participants were also told how they would be evaluated for each task (e.g., “For commander queries, your performance will be evaluated on the time it takes you to respond as well as the accuracy of your response”).
Once participants had been trained on the mission and the UV capabilities, they were trained on the functionality shared across IMPACT and Baseline including map movement, map decluttering, chat, vehicle dashboard, vehicle summary panel, media manager, and help. Participants were then given a high-level overview of the autonomous systems, including the CCA, the IA, the Plan Monitor, the Task Manager, and the voice recognition system.

Next, participants were trained on Baseline, and how to use the system to respond to each possible type of mission event. After a break, participants completed a sixty minute Baseline capstone mission, equivalent in complexity to a low complexity experimental scenario. During the capstone mission the lead experimenter answered any questions the participant had, pointed out any errors the participant made, and suggested better methods for accomplishing tasks.

After a break for lunch, the participant was trained on IMPACT, including the voice recognition system, the Play Calling interface, the Play Workbook, the Active Play Manager, the COA Planner, and the Task Manager. Participants were then given an opportunity to respond to each possible mission event using IMPACT. After a break participants completed a sixty minute IMPACT capstone mission, equivalent in complexity to a low complexity experimental scenario. Just as during Baseline capstone, the lead experimenter answered any questions the participant had, pointed out any errors the participant made, and suggested better methods for accomplishing tasks.
Day 2: Experimental Trials. On the second day, participants completed four sixty minute experimental trials blocked by system. Participants were given refresher training before each block. Refresher training consisted of the participant being asked to respond correctly to chat requests for each RAM, each normal base defense event, each intruder event, each commander query type (time to get a specific vehicle to a specific location, time to get a specific capability to a specific location, time to get a specific task in place, vehicle speed, vehicle altitude, what a vehicle was doing, and what vehicle was doing/had done a specific task), sensor and vehicle failures, environmental events, and ROZs.

Once refresher training was completed participants were given a paper copy of the RAMs they were responsible for conducting in the first trial and given as much time as they needed to develop a plan for executing the RAMs. When the participant was ready, the lead research counted down (“Three, two, one, GO!”) and the mission began. During the mission the lead researcher sat alongside the participant and observed his or her actions. The lead experimenter did not intercede unless the participant encountered a software bug or to prevent a participant from crashing the system due to a known bug. Both the lead experimenter and the TOC operator recorded how well participants did on each mission task to supplement Fusion’s data logs (see APPENDIX H for a sample score sheet). A software tool called Camtasia was used to record the Sandbox screen and well as all voice commands and radio calls the participant made during the mission.
Dependent Measures

Subjective Measures. After each trial participants completed the NASA-TLX. After each block participants completed a tool specific overall questionnaire (see APPENDIX I) and a tool specific usability scale (see APPENDIX J). After the participant had completed both blocks, he or she filled out a questionnaire comparing IMPACT to Baseline across mission tasks (see APPENDIX K).

Objective Measures. Performance data for each type of mission event (RAMs, Normal Base Defense Events, Intruder Events, Vehicle Failures and Environmental Events, and Commander Queries) were collected. For RAMs, participants were scored on how many RAMs they accomplished correctly (i.e., met all the constraints for) during the course of the mission. For Vehicle Failures and Environmental Events, participants were scored on how many they responded to correctly. Both accuracy and response time (the time the query was sent to the participant’s chat window until the participant replied via chat) data was collected for Commander Queries.

For Normal Base Defense Events and Intruder Events both outcome (i.e., was the response to the event accomplished correctly) and process (i.e., did the participant select the correct location/target, correct play, the best vehicle, and meet the event’s constraints) data were collected. For example, imagine that a participant, in response to a task to provide an escort for Convoy Kilo before Convoy Kilo left the gate at 22:00, called an overwatch for Convoy Kilo that wasn’t in place until 23:00. In this case the outcome
score would be 0 because the participant called the wrong play (an overwatch instead of an escort) and was late getting the play in place. However, the process score would be 0.5 because the participant called the play on the right target (Convoy Kilo) and picked an appropriate UV (each component of the process score was equally weighted).

Analyzing the data in this fashion provided information that was both mission relevant (i.e., did the mission get accomplished?) and diagnostic of where the process may have broken down (e.g., participants often failed to respond correctly to a specific event because they had trouble identifying the best vehicle to use).

Response time was not analyzed for mission events because direct comparisons between conditions were not possible. For example, in the Baseline Low Complexity scenario a participant may not have even attempted to respond to a particular event, while responding to the same event in the IMPACT Low Complexity scenario. Instead, the time and number of mouse clicks from when a participant began to call a play (e.g., click a play icon in the Play Calling Tile, click a play icon in the UxAS Wizard) until the play was executed (e.g., hit the check mark on the Play Calling, hit the check mark on the UxAS Wizard) was analyzed.
CHAPTER VI. Results

IMPACT vs. Baseline Subjective Measures

Overall Ratings. Participants used a 5-point Likert scale (ranging from No Aid to Great Aid) to rate IMPACT and Baseline across three measures: potential value to future UV operations, ability to aid operator workload in future UV operations, and ability to aid situation awareness in future UV operations. The data was analyzed using a paired samples t-test. IMPACT was rated significantly higher than Baseline for both potential value to future UV operations, \( t(7) = 3.99, p = .005, d = 1.99^{16} \) and ability to aid workload \( t(7) = 5.35, p = .001, d = 5.86 \) (Figure 39). In fact, all eight participants gave IMPACT the highest possible rating when asked about IMPACT’s potential value to future UV operations and seven out of eight participants gave IMPACT the highest possible score when asked about IMPACT’s ability to aid workload in future UV operations. No significant difference was found for the ability to aid situation awareness, \( t(7) = 1.49, p = .18, d = 0.54 \) (Figure 39).

\[16\] Because there was no variance in the IMPACT rating (i.e., all eight participants gave IMPACT the highest possible rating), the statistics were calculated by pooling the variance.
Figure 39. Comparison of Overall Ratings for IMPACT and Baseline.

**System Usability Scale.** The overall usability of each tool was assessed using the System Usability Scale (SUS) (Brooke, 1996). The SUS asks participants to evaluate 10 items related to system usability using a 5 point Likert scale ranging from Strongly Agree to Strongly Disagree, and these 10 items contribute to an overall SUS score. Participants rated IMPACT higher than Baseline on every single SUS item (see Figure 40—note that odd numbered questions are positive statements and even numbered questions are negative statements, therefore high scores are better for the odd numbered questions and low scores better for the even numbered questions). The overall SUS scores were
compared using a paired samples t-test. IMPACT’s overall SUS score was significantly higher than Baseline’s overall SUS score, \( t(7) = 2.73, p = .03, d = 0.97 \) (see Figure 41).

![Figure 40. Comparison of System Usability Scale (Brooke, 1996)](image)

<table>
<thead>
<tr>
<th>1. I think that I would like to use this system frequently.</th>
<th><strong>Strongly Disagree</strong></th>
<th><strong>Strongly Agree</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>2. I found the system unnecessarily complex.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>3. I thought the system was easy to use.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>4. I think that I would need the support of a technical person to be able to use this system.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>5. I found the various functions in this system were well integrated.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>6. I thought there was too much inconsistency in this system.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>7. I would imagine that most people would learn to use this system very quickly.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>8. I found the system very cumbersome to use.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>9. I felt very confident using the system.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>10. I needed to learn a lot of things before I could get going with this system.</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
</tbody>
</table>

Figure 41. Comparison of SUS Score for IMPACT and Baseline.
IMPACT vs. Baseline. After the experimental trials were completed, for each mission task, participants were asked to rate whether they performed the task better with IMPACT or better with Baseline. Participants rated their performance as better with IMPACT as compared to Baseline for every single mission task (see Figure 42 – gray dots represent the mean rating for each task).

Figure 42. IMPACT vs. Baseline Ratings for Mission Tasks.

Participants were also given the opportunity to comment on the differences between IMPACT and Baseline. Two participants elected not to comment. Of the remaining six participants, five (participants 1, 4, 5, 6, and 8) were very positive about
IMPACT as compared to Baseline (Table 6). A single participant (participant 3) gave IMPACT a mixed review stating that although his or her performance was better with IMPACT, he or she had better situation awareness and confidence with Baseline.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“IMPACT workload (or ability to absorb) is incredible over baseline. S4 [a high complexity scenario], could never be executed [with baseline]. Workload, automation, and automation-assist were light years ahead of baseline.”</td>
</tr>
<tr>
<td>3</td>
<td>“Although I think I performed better, and tasks were overall easier to accomplish w/Impact, I had greater confidence and situational awareness using baseline system”</td>
</tr>
<tr>
<td>4</td>
<td>“Impact by far outperformed baseline. Just more user friendly than baseline. Very simple to understand.”</td>
</tr>
<tr>
<td>5</td>
<td>“Overall IMPACT is much better to use. In high tempo environments with many things happening at once, IMPACT is the standard that should be used all the time.”</td>
</tr>
<tr>
<td>6</td>
<td>“IMPACT is so superior in how it assists the operator, eliminating much of the ‘planning/thinking’ aspects and confusion, compared to the Baseline, it's ridiculous. The reduction in stress and time required to perform almost every task, using IMPACT vs. the Baseline is amazing, and a welcome, mental relief!”</td>
</tr>
<tr>
<td>8</td>
<td>“Tasks performed much faster with IMPACT. Easier to keep track of unfinished tasks w/ IMPACT. Easier to plan ahead w/ IMPACT. Tasks compounded in Baseline.”</td>
</tr>
</tbody>
</table>

NASA-TLX. Participants completed the NASA-TLX to assess their perceived workload after each experimental trial. Data were analyzed with a repeated measures Analysis of Variance (ANOVA) and shown in Figure 43. No significant interaction between tool and complexity was found, \((F(1,7) = 2.57, p = .15, \eta_p^2 = .27)\). The results indicated a main effect of complexity \((F(1,7) = 17.06, p = .004, \eta_p^2 = .71)\), with participants rating workload lower in the Low Complexity condition \((M = 36.72)\) than in the High.
Complexity condition \( (M = 58.54) \). The results did not indicate a main effect of tool \( (F(1,7) = 4.08, p = .08, \eta^2_p = .27, \text{Impact } M = 43.28, \text{Baseline } M = 51.98) \).

Figure 43. NASA-TLX Results.

**IMPACT vs. Baseline Objective Measures**

**RAMs.** For RAMs, participants were scored on how many RAMs they accomplished correctly (i.e., met all the constraints for) during the course of the mission. The percentage of RAMs completed was calculated by dividing the number of RAMs successfully completed during a mission by the total number of RAMs that participants were given at the start of the mission. Data were analyzed with a repeated measures
analysis of variance (ANOVA) and the results are shown in Figure 44. No significant interaction between tool and complexity was found, \( (F(1,7) = 0.14, p = .72, \eta_p^2 = .02) \). The results indicated a main effect of tool \( (F(1,7) = 7.68, p = .03, \eta_p^2 = .52) \), with participants completing more RAMs in the IMPACT condition \( (M = 62.40\%) \) as compared to the Baseline \( (M = 40.97\%) \) condition. The results also indicated a main effect of complexity \( (F(1,7) = 28.45, p = .001, \eta_p^2 = .80) \) with participants completing more RAMs in the low complexity condition \( (M = 58.93\%) \) as compared to the high complexity condition \( (M = 44.44\%) \).

![Figure 44. Percentage of RAMs Completed Correctly.](image)

**Normal Base Defense Events.** Both outcome and process data were collected for normal base defense events and analyzed with repeated measures ANOVAs. The outcome score was calculated by taking the number of normal base defense events the
participant responded to correctly divided by the number of normal base defense events. To compute the process score, each participant’s response to normal base defense events was also graded on four subcomponents: (1) did the participant select the correct location/target; (2) did the participant select the correct play; (3) did the participant select the best UV; and (4) did the participant meet the event’s constraints. Each component was weighted equally and the average of the four components was used as the process score. For example, imagine if a participant, in response to a task to investigate a suspicious vehicle at Point Alpha, called an air inspect at Point Alpha using the fastest UV, but didn’t leave the UV at Point Alpha until the all clear was given. In this case the outcome score would be 0 because the participant failed to keep the UV in place until the all clear. However, the process score would be 0.75 because the participant called the play at the right location (Point Alpha), called the correct play (Air Inspect) and picked the fastest UV.

Results for outcome score are shown in Figure 45. No significant interaction between tool and complexity was found ($F(1,7) = 0.50, p = .50, \eta^2_p = .07$). The results did not indicate a main effect of tool ($F(1,7) = 5.49, p = .051, \eta^2_p = .44$, IMPACT $M = 73.92\%$, Baseline $M = 50.89\%$). The results also failed to indicate a main effect of complexity ($F(1,7) = 3.77, p = .09, \eta^2_p = .35$, low complexity $M = 71.25\%$, high complexity $M = 53.57\%$).
Results for process score are shown in Figure 46. No significant interaction between tool and complexity was found ($F(1,7) = 2.97, p = .13, \eta_p^2 = .30$). The results indicated a main effect of tool ($F(1,7) = 7.57, p = .03, \eta_p^2 = .52$), with participants performing better in the IMPACT condition ($M = 88.84\%$) as compared to the Baseline ($M = 76.69\%$) condition. The results also indicated a main effect of complexity ($F(1,7) = 16.09, p = .005, \eta_p^2 = .70$) with participants performing better in the low complexity condition ($M = 92.81\%$) as compared to the high complexity condition ($M = 72.71\%$).
The overall pattern of the outcome and process measures were similar, with participants scoring higher in the IMPACT condition as compared to the Baseline condition and in the low complexity condition as compared to the high complexity condition. Failing to reach the alpha level of .05 for the outcome measure may be more indicative of the small number of participants and high level of variance imposed by the all or nothing scoring scheme than it is an indicator of a lack of performance difference between conditions.

*Intruder Events.* Both outcome and process measures were collected for Intruder Events, using the same process described above for normal base defense events. Results for outcome score are shown in Figure 47. No significant interaction between tool and
complexity was found ($F(1,7) = 0.07, p = .90, \eta^2_p = .001$). The results did not indicate a main effect of tool ($F(1,7) = 3.21, p = .12, \eta^2_p = .31$, IMPACT $M = 69.87\%$, Baseline $M = 55.58\%$). The results did indicate a main effect of complexity ($F(1,7) = 9.91, p = .02, \eta^2_p = .59$, with participants performing better in the low complexity condition ($M = 75\%$) than in the high complexity condition ($M = 50.45\%$).

![Figure 47. Intruder Events Outcome Measure.](image)

Results for process score are shown in Figure 48. No significant interaction between tool and complexity was found ($F(1,7) = 3.21, p = .12, \eta^2_p = .31$). The results indicated a main effect of tool ($F(1,7) = 33.90, p = .0007, \eta^2_p = .83$), with participants performing better in the IMPACT condition ($M = 88.54\%$) as compared to the Baseline $M = 50.45\%$).
(M = 69.22%) condition. The results also indicated a main effect of complexity (F(1,7) = 30.24, \( p = .0009, \eta_p^2 = .81 \)) with participants performing better in the low complexity condition (M = 89.32%) as compared to the high complexity condition (M = 68.43%).

![Figure 48. Intruder Event Process Measure.](image)

As with the Normal Base Defense Events, the outcome and process scores followed the same general pattern, with participants performing better with IMPACT as compared to Baseline and better in the low complexity condition as compared to the high complexity condition. Again, failing to reach the alpha level of .05 for the outcome measure may be more indicative of the small number of participants and high level of
variance imposed by the all or nothing scoring scheme than it is an indicator of a lack of performance difference between conditions.

*System Failures and Environmental Events.* Participants were scored on how many system failures and environmental events they responded to correctly (i.e., followed the QRC for) during the course of the mission. The percentage of system failures and environmental events responded to correctly was calculated by dividing the number of events responded to correctly during the mission by the total number of events that occurred over the duration of the mission. Data were analyzed with a repeated measures analysis of variance (ANOVA) and the results are shown in Figure 49.

No significant interaction between tool and complexity was found, (F(1,7) = 0.25, p = .63, $\eta_p^2 = .04$). The results did not indicate a main effect of tool (F(1,7) = 0.07, p = .80, $\eta_p^2 = .01$), IMPACT $M = 76.22\%$, Baseline $M = 74.36\%$. The results did indicate a main effect of complexity (F(1,7) = 17.20, p = .004, $\eta_p^2 = .71$) with participants responding correctly to more system failures and environmental events in the low complexity condition ($M = 86.46\%$) as compared to the high complexity condition ($M = 64.11\%$).
These results match well with participant’s subjective ratings. See for instance, Figure 42—participants rated their performance on responding to system failures and environmental events as better with IMPACT, but less so than other tasks. These results most likely reflect the lack of difference in functionality between IMPACT and Baseline when responding to system failures or environmental events. For example, if dense smoke was reported in an area, the response the participant was required to make—switching any affected UV’s sensor to IR—was accomplished in the same manner in both IMPACT and Baseline.
Commander Queries. Both accuracy and response time data were collected for commander queries and analyzed with a repeated measures ANOVA. Accuracy data is shown in Figure 50. For accuracy, no significant interaction between tool and complexity was found, \((F(1,7) = 0.005, p = .95, \eta_p^2 = .0007)\). The results indicated a significant main effect of complexity \((F(1,7) = 17.39, p = .004, \eta_p^2 = .71)\), with participants answering commander queries more accurately in the low complexity condition \((M = 91.09\%)\) than the high complexity condition \((M = 76.68\%)\). There was not a significant main effect for tool, \(F(1,7) = 3.33, p = .11, \eta_p^2 = .32\).

![Figure 50. Commander Query Accuracy.](image)

For response time (shown in Figure 51) a significant interaction \((F(1,7) = 6.39, p = .04, \eta_p^2 = .48)\) was found. In the low complexity condition participants were faster at
answering queries with IMPACT as compare to Baseline. However, in the high complexity condition participants answered commander queries faster with Baseline as compared to IMPACT.

Figure 51. Commander Query Response Time.

At first, the results of this analysis were perplexing—why were participants faster at answering queries with Baseline in the high complexity condition? Upon examining the video recordings of the high complexity trials an interesting pattern of behavior emerged. In the high complexity scenarios, some participants using IMPACT would often set commander queries aside in order to focus on higher priority tasks (i.e., normal base defense events or intruder events) and return to the query later. In contrast, in the
Baseline condition, these participants would immediately answer the query instead of responding to the higher priority tasking. It appeared as if some participants in the Baseline condition were relieved when a commander’s query came in asking them, “What’s TR-22’s Speed?” in a high complexity scenario because it was an easy task that they knew how to answer. IMPACT, on the other hand, seemed to help participants prioritize tasks and enabled them to have discretionary control. The performance data supports this hypothesis. In Figure 52, each participant’s average process score for normal base defense events and intruder events is mapped on the y axis, while response time to commander queries is mapped on the x axis. Baseline data is coded blue and IMPACT data is coded red. Note, that of the Top 10 average process scores, 8 of them occur when the participant was using IMPACT. Also note that though three participants (1, 2, and 4) had noticeably slower mean query response times with IMPACT, all three had higher average process scores with IMPACT.
Figure 52. Commander Query Response Time vs. Process Score.

**Time/Number of Clicks to Call Plays.** The time from when a participant began to call a play (e.g., click a play icon in the Play Calling Tile, click a play icon in the UxAS Wizard) until the play was executed (e.g., hit the check mark on the Play Calling, hit the check mark on the UxAS Wizard) was analyzed using a repeated-measures ANOVA (see Figure 53). No significant interaction was found ($F(1,7) = 1.00, p = .35, \eta^2_p = .13$).

Additionally there was neither a significant main effect of tool ($F(1,7) = 0.14, p = .72, \eta^2_p = .02$, Impact $M = 18.67$, Baseline $M = 19.3$) nor a significant main effect of complexity ($F(1,7) = 0.22, p = .65, \eta^2_p = .03$, low complexity $M = 18.53$, high complexity $M = 19.44$).
These results were initially surprising because it was expected that participants would be able to call plays more quickly using IMPACT. However, a key point to make is that in Baseline, participants had to assign all 12 UVs to a different play when switching patrol states. These play calls become very rote (i.e., participants did not have to think about which vehicle to send where), and participants were able to quickly execute these plays using the UxAS Wizard, thus lowering the overall mean time for Baseline play calls. An attempt was made to identify and separate these plays, but it was determined that it would be very difficult and time consuming to identify these plays from the data logs.
However, the number of clicks participants were required to make from play initiation to play execution were also analyzed using a repeated-measures ANOVA (see Figure 54). Again, no significant interaction was found ($F(1,7) = 1.02, p = .35, \eta^2_p = .13$). However, the results did indicate a significant main effect of tool ($F(1,7) = 390.33, p < .0001, \eta^2_p = .98$) with participants making a smaller number of clicks to execute plays with IMPACT ($M = 2.09$) than Baseline ($M = 4.95$). The results did not indicate a main effect of complexity ($F(1,7) = 0.005, p = .95, \eta^2_p = .0007$, low complexity $M = 3.52$, high complexity $M = 3.52$).

![Figure 54. Mean Number of Clicks from Play Initiation to Play Execution.](image)
Thus, even though there was not a significant difference between the time it took participants to execute plays with IMPACT as compared to Baseline, participants were able to execute plays using a significantly less number of mouse clicks with IMPACT.

**IMPACT vs. Baseline Summary.**

Table 7 provides a summary of all the analyses described above, for both subjective and objective measures. Participants rated IMPACT significantly better than Baseline in terms of its perceived value to future UV operations as well as its ability to aid workload. In fact, every single participant gave IMPACT the highest possible score for potential value, and all but one participant gave IMPACT the highest possible score for its ability to aid workload. IMPACT was not rated significantly better than Baseline in terms of its ability to aid situation awareness, a result echoed by Participant 3’s comment that his or her situation awareness was better with Baseline. In general, participants expressed that map clutter lowered their situation awareness in both conditions.

Participants also performed significantly better with IMPACT across multiple performance measures including the number of RAMs completed and the process score for both normal base defense events and intruder events. Despite failing to reach an alpha level of .05 for the normal base defense events and intruder event outcome measures, the overall pattern of the results indicates a similar trend as the process measures.
Finally, the results indicated that participants performed better in the low complexity scenario as compared to the high complexity scenario across almost all performance measures.

Table 7. IMPACT vs. Baseline Summary.

<table>
<thead>
<tr>
<th>Measure</th>
<th>IMPACT &gt; Baseline</th>
<th>IMPACT &gt; Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subjective Measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potential Value</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Aid Workload</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Aid Situation Awareness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System Usability Scale</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>NASA-TLX</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Objective Measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rams Completed Correctly</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>Normal Base Defense Outcome Measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Base Defense Process Measure</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>Intruder Events Outcome Measure</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Intruder Events Process Measures</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Response to System Failures/Environmental Events</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Commander Query Accuracy</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Commander Query Response Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time to Call Plays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Clicks to Call Plays</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* significant at .01
** significant at .05
Qualitative Results

One of the benefits of having a complex synthetic task environment was it brought to the fore problems that would have never arisen in a simple accuracy/response time study. Participants often got into situations, took actions, or used the autonomous system in ways that the experiments didn’t anticipate. In this section, multiple use cases are presented that showcase how the synthetic task environment enabled the experimental team to identify limitations of IMPACT and design improvements.

Physical Distance vs. Estimated Time Enroute. A common error made in the Baseline condition is shown in Figure 55. The participant was tasked to get eyes on a suspicious vehicle as soon as possible and called an Air Inspect at the target’s location with FN-42. However, though FN-42 was physically closer to the target’s location, FN-40 could have gotten to the location in less time because it is much faster than FN-42.
In the IMPACT condition, the IA would recommend using FN-40 in the situation depicted in Figure 55. However, unless participants opened up the COA planner, they were left to wonder why the IA had recommended a UV that was further away. Based on this result, a new interface widget (see Figure 56) was designed to show how close vehicles are to a location in terms of time rather than space. With this representation vehicle distance from the target location (represented by the black dot in the middle) is determined by the time it would take for the vehicle to arrive to the point as opposed to the physical distance the vehicle is away. The range rings represent time increments—1 minute, 3 minutes, and 5 minutes.
Figure 56. Time to Arrive.

*Short-Term Optimization Leads to Long Term Suboptimization.* The IA would always try and recommend the best UV to complete a play, but would not consider its recommendation within the context of other ongoing mission tasks. This led to several instances during the experiment in which each task had been individually optimized, but the overall response was not. For example, in one situation a participant was providing overwatch for a convoy with FN-41 when a mortar fire was reported. Following the QRC, the participant called a play to get eyes-on the mortar fire site as soon as possible, and accepted the IA recommendation to use FN-40. The next step in the QRC was to search for the individual who fired the mortar with a UAV optimized for tracking. In this instance, the two UAVs with the best tracking capabilities—FN-40 and FN-41—were already occupied, and thus, excluded from the IA’s optimal recommendation. Thus, the IA recommended that the participant use FN-42 to search for the individual who fired the
mortar, as FN-42 was the next best tracker available, which the participant accepted. A better solution would have been to inspect the mortar site with FN-42 and search for the individual who fired the mortar with FN-40. To better assist the operator in these types of situations the IA is being redesigned to consider the best UV to use in the context of all ongoing tasks. Additionally, new user interface concepts are being designed to allow the human operator to ask the IA to critique the current task assignments and search for possible better allocations.

*Hammer of Thor.* Recall from Chapter III that IMPACT provided a method for operators to tell the autonomy to cancel whatever plays needed to be cancelled in order to respond to a high priority event. During the experiment, a single participant used this “Hammer of Thor” feature and the results were illuminating. To set the stage, the participant was in the middle of an intruder event in a high complexity scenario. FN-43, TR-22, and TR-23 were escorting a manned unit outside the base perimeter, FN-40 and TR-20 were responding to the intruder event, TR-21 was conducting a Show of Force, and the rest of the UVs were in a highly mobile patrol state. At this point the participant received a chat message about a building alarm, and in response, called a Ground Inspect play at the building. However, all of the UGVs were currently on tasks, so the play could not be executed immediately. At this point the participant hit the “Hammer of Thor” which instructed the IA to assign a UGV to conduct the Ground Inspect regardless of any other tasks. By default, the IA decided which UGV to send based on time, which happened to be TR-22. When TR-22 was reassigned to conduct the ground inspect, the escort play
TR-22 had been conducting for the manned unit outside the base perimeter was paused. Unfortunately, this also caused the other two vehicles on the escort, FN-43 and TR-23, to stop the escort as well.

When this occurred, the participant froze when he realized the escort had been paused to fulfill the ground inspect play at the building alarm. The participant tried to recall the escort play, but obviously there were no longer enough assets to successfully fill the escort play request. At this point the participant exclaimed, “Okay I jacked that up pretty bad. I don’t know how to unjack that so I’ll just have to come back to it”.

To remedy this situation, interface concepts are being designed that allow the human operator to see which plays will be cancelled if the “Hammer of Thor” feature is used. Additionally, the concept of play priority is slowly being added in beginning with RAM plays. When a play is called, it can now be designated as a RAM, which informs the autonomous system components that the play is low priority (and should be cancelled when an intruder threat occurs). If this feature had been implemented in the above scenario, the autonomous system would have cancelled TR-21’s show of force and used it to investigate the building alarm, thus keeping the escort intact.
CHAPTER VII. Discussion

The hypothesis that participants would both prefer and perform better with IMPACT’s HAI as compared to Baseline was supported. Participants preferred IMPACT as compared to Baseline on multiple subjective measures including usability, perceived value to future UV operations, and ability to aid workload. Participants also performed better with IMPACT as compared to Baseline on multiple objective measures including number of RAMs completed and the process score for both normal base defense events and intruder events.

The hypothesis that Operator performance would be worse in the high complexity missions as compared to low complexity was supported, with participants performing better in the low complexity missions across almost all performance measures. However, the hypothesis that the performance difference between IMPACT and Baseline would be significantly greater in the high complexity scenarios was not supported. Several factors may account for this including a lack of statistical power due to the small number of participants as well as limiting the experiment to two levels of complexity.

Limitations

This research effort had multiple limitations. First and foremost, this study was limited to a small number of participants due to budgetary, time, and availability constraints. The small number of participants reduced the statistical power of the study and thus increased the susceptibility to a type II error. For example, the outcome
measure difference between IMPACT and Baseline for both Normal Base Defense Events and Intruder Events was not significantly different despite a seemingly large difference in the means.

At the beginning of this research effort it was determined that the advantages of using participants with real world experience would outweigh the negatives. One of the negatives was the lack of time to train participants. In the operational world, a Warfighter would have far more time to learn the tool—two weeks instead of a single day—before needing to use the tool in a real-world mission. Unfortunately, it was extremely difficult to find active-duty participants who could donate two days of their time let alone two weeks. It is unreasonable to expect that participants could expertly wield all of IMPACT’s functionality after a single day of training. In fact, certain results, such as participants not using the voice recognition system to call plays, may be directly tied to a lack of training. Valuable future research would include replicating this experiment with extensively trained participants, and it would be valuable even if these participants were not military professionals.

Due to software development that continued up until the eleventh hour, it was impossible to thoroughly vet the experimental scenarios before running participants. Thus, the low complexity scenario may have been a bit too easy and the high complexity scenario may have been a bit too difficult. Additionally, it was difficult to tease out the value of individual autonomous components (e.g., the CCA, the IA, etc.). The original
experimental design consisted of multiple tool levels (adding in autonomous components) and multiple levels of complexity. However, due to software and budget constraints this was modified to a 2X2. If this study were to be replicated, it could benefit from a few additional levels of each independent variable (i.e., gradually adding in automation capabilities and having four scenarios of gradually increasing complexity).

**Scientific Contribution**

As noted in Chapter III, the majority of EID interfaces have focused on providing support for manual control tasks, which typically involve skill-based behaviors (Borst, Flach, & Ellerbroek, 2015). The results of this experiment provide evidence that ecological interface design principles can be used to generate user interface concepts that not only support skill-based behaviors, but also rule-based and knowledge-based behaviors. Additionally, this study demonstrates that ecological interface design principles can be generalized to domains other than primary flight displays and nuclear power process control.

This research also demonstrates the power of conducting a study in a synthetic task environment. Placing the operator and autonomy in a simulated mission and having them collaborate to perform tasks derived from real word incidents provided invaluable feedback on where the team struggled as well as how IMPACT could be improved.

Finally, this research has contributed to the ongoing debate within the research community about how to design effective autonomous systems. In fact, the first chapter
of this dissertation has been published in the journal *She-Ji: The Journal of Design, Economics, and Innovation*, and sparked a lively discussion and official replies from Derek Miller, Hugh Dubberly and Paul Pangaro, and Susu Nousala. Furthermore, many of the smaller studies conducted in the course of this research have also been published (Behymer, Mersch, Ruff, Calhoun, & Spriggs, 2015; Behymer, Ruff, Mersch, Calhoun, & Spriggs, 2015; Mersch, Behymer, Calhoun, Ruff, & Dewey, 2016; Calhoun, Ruff, Behymer, & Mersch, 2016).

### Future Research

In addition to the future research options discussed in the limitations section, a significant challenge facing the design team is determining the number of alternative plans to present to the operator. One approach is to present a single solution to the operator, the solution that the autonomy has determined to be the best. Another approach that has been advocated is modeling to generate alternatives (MGA) The MGA approach focuses on generating a small set of alternatives that are “good” in terms of achieving the operator’s goal but different in respect to the relevant parameters of the solution space (Brill, Flach, Hopkins, and Ranjithan, 1990). The MGA approach suggests the system should generate a set of options that it believes can achieve the commander’s intent (i.e., whatever presets are specified) but vary in other parameters. For example, if the operator asks for getting a UAV to Point Charlie as soon as possible, Plan A could also minimize fuel use, Plan B could maximize stealth, and Plan C could minimize impact on response
time to the flight line. Future research should examine the utility of applying the MGA approach to IMPACT’s COA Planner.

Several key components of current and envisioned UV operations were not examined in this research, including collision avoidance and deconfliction, Multi Operator-Multi UV (MOMU) operations in which multiple human operators share a number of UVs, and Manned-Unmanned Teaming (MUM-T) operations in which UV operators work with manned units to accomplish tasks. Future research should investigate how IMPACT can be extended to these mission sets and operational concerns.

**Conclusion**

Chapter I discussed the limitations of the substitution approach—replacing humans with autonomy—and suggested an alternative approach to autonomous system design. A collaborative systems approach was proposed, that would enable the operator and the autonomy to work together, capitalizing on each other’s strengths, to accomplish mission tasks. Rather than rehash the arguments against the substitution approach described in Chapter I, this summary will briefly touch on another line of research within the autonomous systems research community—mutual understanding—and discuss its limitations.

Throughout its history, psychology has vacillated between focusing primarily on the mind and focusing primarily on behavior. Titchener’s structuralism was followed by Watson and Skinner’s behaviorism, which in turn gave way to information processing.
This legacy has significantly influenced today’s cognitive psychologists, who are wont to think that cognition is about *understanding* the world. An alternative approach suggests that cognition is about *acting* in the world (Rączaszek-Leonardi, 2014). The traditional approach posits that people communicate in order to understand each other. In contrast, Rączaszek-Leonardi proposes that people communicate to accomplish things together.

Many psychologists who study human-autonomy interaction approach their research from the traditional perspective. Hence, much current human-autonomy interaction research is focused on making the autonomy understandable to the user, with a goal of equalizing the mental model of the human operator and the autonomy (Rączaszek-Leonardi, 2014). Unfortunately, this paradigm takes the focus away from the work that the human-autonomy team is trying to accomplish.

Rączaszek-Leonardi (2014) claims that though people might never understand each other, they can do something far more important: work together to accomplish things. Rączaszek-Leonardi’s approach echoes Cooke’s work on team cognition. While the traditional approach to team cognition assumes that knowledge homogeneity (i.e., all team members have the same knowledge) is a good thing, Cooke (2015) argues that team knowledge heterogeneity is essential to team success in dynamic environments. For Cooke, team cognition *is not* team member shared knowledge, but rather *is* team interaction.

This applies to human-autonomy teams as well. Recall Figure 1, in which both humans and agents are limited in their ability to perceive the state of the world as well as
the control actions they can take. However, as Rączaszek-Leonardi states, if the teammates’ knowledge, skills, perspectives, and control capabilities differ they will be much better equipped to satisfy the demands of Ashby’s law of requisite variety. Thus, HAI s are not needed to help humans and autonomy understand each other; instead, HAI s are needed that enable humans and autonomy to work together to accomplish things.

This has important implications for HAI design. Starting with the goal of designing an HAI that allow humans and autonomy to understand each other runs the risk of the HAI being completely irrelevant to the work. As Rączaszek-Leonardi (2014) states, as long as each actor’s moves are proper and relevant in a given situation, understanding the actor’s mental state is not only unimportant, but also may prove costly. Providing the ability for the human and autonomy to understand each other is important when a surprise occurs—when one of the actor’s makes a move that is improper or irrelevant for the situation. However, the goal of an HAI should always be focused on accomplishing the work, not creating mutual understanding for the sake of mutual understanding.

To design a work-focused HAI it is vital to understand the work domain. To this end, Chapter II focused on gaining an understanding of the base defense domain using cognitive task analysis methodologies. Interviews were conducted with base defense operators, search and rescue personnel, and UVs operators in order to understand how UVs could support base defense missions. The results of the CTA informed the
design of the autonomous systems, the HAI, and the experimental scenario used to validate IMPACT.

Chapter III described the development of IMPACT’s HAI, which enables a single operator to work with several autonomous components to manage multiple UVs in the context of a base defense mission. IMPACT’s interfaces were designed to support skill-based, rule-based, and knowledge-based interactions and to support direct perception and direct manipulation. Though these EID principles certainly guided and shaped the development of IMPACT’s HAI, the role of serendipity and art also played a role. For example, video games influenced the design of the Play Calling Interface and a happenstance encounter with a parallel coordinates plot influenced the design of the COA planner.

It would have been easy to develop a laboratory task to examine how IMPACT improved participant’s speed and accuracy as compared to a Baseline condition. For example, participants could have been instructed to call plays in response to a series of independent events. Though this type of study may have yielded statistically rigorous and significant data, it would have lacked pragmatic utility. Perhaps this type of research is at the root of psychology’s replication problem (Open Science Collaboration, 2015). Regardless, studying the phenomena of interest with an entire focus on experimental control runs the risk of driving away the very phenomena one is interested in.

To this end, this research utilized a synthetic task environment, in which the demands of a real world base defense mission were simulated to the fullest extent within
time and budgetary constraints. Eight participants with relevant military experience (UV operations and base defense) were given free rein to command multiple UVs during an hour-long base defense mission. The experimental design made data analysis particularly challenging, however it also provided the enormous benefit of seeing how the human-autonomy team responded to situations that even the experimental team hadn’t envisioned.

Still, traditional subjective and objective performance measures were, to the extent possible, collected to compare operator performance using IMPACT as compared to a Baseline condition that was considered state-of-the-art at the beginning of the IMPACT project. Participants both preferred and performed better with IMPACT as compared to Baseline on multiple performance measures. Participant feedback as well as use cases directly led to improvements to IMPACT’s HAI. For example, after participants discussed the value that adding temporal information to the Active Play Manager would add to their ability to fulfill their mission, this feature was immediately added. Future work will seek to expand this temporal information into a timeline display that will enable operators to better plan out future tasks and conduct retrospection if a UV performed tasks independently while out of communication range.
AFTERWORD

“But I always want to know the things one shouldn’t do.” “So as to do them?” asked her aunt. “So as to choose,” said Isabel. – (James, 1881)

When I left Thomas More College in 2002 to attend the Human Factors Psychology program at Wright State University I felt I knew a lot about psychology. My experience in the program, specifically in my advisor Dr. John Flach’s classes, quickly made me realize how little I, and in fact anyone, truly knew. I had left undergrad believing that information processing was right. John, in the words of Edith Wharton (1934), is “supremely gifted as an awakener, and no thoughtful mind can recall without a thrill the notes of the first voice which has called it out of its morning dream”. John woke me up from my information processing dream. Or to put it into more recent pop culture terms, John offered me the red pill and I took it.

One night, early in my graduate career a man named Dennis Allen wandered into the lab asking if there were any cognitive psychologists around. I answered, “Buddy, at this time of night I’m about the best you are going to do.” This encounter led to my first industry job as a research psychologist with JXT Applications. Soon, I was doing my best to put the theories and methods I had been exposed to in the classroom to use. Six months prior I had never heard of cognitive task analysis. Now I was trying to conduct CTA interviews with B1 bomber operators. These experiences allowed me to see the wide chasm between academia and the real world. Instead of having years to conduct perfectly controlled experiments, I had less than six months to complete an entire project.
and demonstrate enough practical utility to receive additional funding. The word satisficing quickly entered my vocabulary, and I quickly became disillusioned with academia.

Thankfully, John was supportive and pointed me in the direction of William James, Robert Pirsig, John Dewey, and ecological psychology, which finally allowed me to connect the worlds of industry and academia. This dissertation is the culmination of these efforts and my collaboration with John over the last 15 years. Any mistakes within this document are my own. I’d like to close with the words of Edward C. Tolman, as I can’t think of a better statement to drop the mic with:

“Since all the sciences, and especially psychology, are still immersed in such tremendous realms of the uncertain and the unknown, the best that any individual scientist, especially any psychologist, can do seems to be to follow his own gleam and his own bent, however inadequate they may be. In fact, I suppose that actually this is what we all do. In the end, the only sure criterion is to have fun. And I have had fun.”

(Tolman, 1959).

-Kyle Behymer
Waynesville OH, January 2017
CHAPTER VIII. References


148


CHAPTER IX. Appendices

APPENDIX A. Acronym List.
APPENDIX B. Random Anti-Terror Measures Lists.
APPENDIX C. Sample Sensor Operator Script.
APPENDIX D. Sample TOC Operator Script.
APPENDIX E. Background Questionnaire.
APPENDIX F. Base Map.
APPENDIX G. UV Comparison Chart.
APPENDIX H. Sample Scoresheet.
APPENDIX I. Overall System Questionnaire.
APPENDIX J. System Usability Scale (Brooke, 1996).
APPENDIX K. IMPACT vs. Baseline Questionnaire.
APPENDIX A. Acronym List.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>ASD R&amp;E</td>
<td>Assistant Secretary of Defense for Research and Engineering</td>
</tr>
<tr>
<td>BDOC</td>
<td>Base Defense Operations Center</td>
</tr>
<tr>
<td>CAP</td>
<td>Civil Air Patrol</td>
</tr>
<tr>
<td>CCA</td>
<td>Cooperative Control Algorithm</td>
</tr>
<tr>
<td>CDO</td>
<td>Cognitive Domain Ontology</td>
</tr>
<tr>
<td>COA</td>
<td>Course of Action</td>
</tr>
<tr>
<td>CTA</td>
<td>Cognitive Task Analysis</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>DCGS</td>
<td>Distributed Common Ground System</td>
</tr>
<tr>
<td>DoD</td>
<td>Department of Defense</td>
</tr>
<tr>
<td>EID</td>
<td>Ecological Interface Design</td>
</tr>
<tr>
<td>EO</td>
<td>Electro-optical</td>
</tr>
<tr>
<td>ETE</td>
<td>Estimated Time Enroute</td>
</tr>
<tr>
<td>HAI</td>
<td>Human Autonomy Interface</td>
</tr>
<tr>
<td>HAL</td>
<td>Human Autonomy Laboratory</td>
</tr>
<tr>
<td>HL2</td>
<td>Half-Life 2</td>
</tr>
<tr>
<td>HUMINT</td>
<td>Human Intelligence</td>
</tr>
<tr>
<td>IA</td>
<td>Intelligent Agent</td>
</tr>
<tr>
<td>IMPACT</td>
<td>Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies</td>
</tr>
<tr>
<td>IR</td>
<td>Infrared</td>
</tr>
<tr>
<td>MDARS</td>
<td>Mobile Detection Assessment and Response System</td>
</tr>
<tr>
<td>MGA</td>
<td>Modeling to Generate Alternatives</td>
</tr>
<tr>
<td>MIC</td>
<td>Mission Intelligence Coordinator</td>
</tr>
<tr>
<td>MMORPG</td>
<td>Massive Multi-player Online Role Playing Game</td>
</tr>
<tr>
<td>MOMU</td>
<td>Multi Operator Multi UV</td>
</tr>
<tr>
<td>MUM-T</td>
<td>Manned-Unmanned Teaming</td>
</tr>
<tr>
<td>NBA</td>
<td>National Basketball Association</td>
</tr>
<tr>
<td>NFL</td>
<td>National Football League</td>
</tr>
<tr>
<td>NICU</td>
<td>Neonatal Intensive Care Unit</td>
</tr>
<tr>
<td>OneSAF</td>
<td>One Semi-Automated Forces</td>
</tr>
<tr>
<td>PoD</td>
<td>Probability of Detection</td>
</tr>
<tr>
<td>QRC</td>
<td>Quick Reaction Checklist</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Anti-terror Measure</td>
</tr>
<tr>
<td>ROZ</td>
<td>Restricted Operating Zone</td>
</tr>
<tr>
<td>SAR</td>
<td>Synthetic Aperture Radar</td>
</tr>
<tr>
<td>SO</td>
<td>Sensor Operator</td>
</tr>
<tr>
<td>Abbr.</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>SRK</td>
<td>Skills, Rules, Knowledge</td>
</tr>
<tr>
<td>UAV</td>
<td>Unmanned Aerial Vehicle</td>
</tr>
<tr>
<td>UGV</td>
<td>Unmanned Ground Vehicle</td>
</tr>
<tr>
<td>USV</td>
<td>Unmanned Surface Vehicle</td>
</tr>
<tr>
<td>UV</td>
<td>Unmanned Vehicle</td>
</tr>
<tr>
<td>UxAS</td>
<td>Unmanned Systems Autonomy Services</td>
</tr>
</tbody>
</table>
APPENDIX B. Random Anti-Terror Measures Lists

RAMs for Scenario A

1. LPOP at Bldg 15 for 5 minutes
2. Imagery of the Ammo Dump every 10 minutes from 0:00:00 to 0:40:00
3. Show of Force (at any gate) for 10 minutes starting at 0:05:00
4. 360 of the Barracks between 0:10:00 and 0:20:00

RAMs for Scenario B

1. Show of Force at Gate 2 for 5 minutes
2. Imagery of the Barracks every 10 minutes from 0:10:00 to 0:50:00
3. LPOP (at any building) for 10 minutes starting at 0:05:00
4. 360 of the Chow Hall between 0:15:00 and 0:25:00

RAMs for Scenario C

1. LPOP at a point inside NAI 4 for 5 minutes at 0:45:00
2. Imagery of the Ammo Dump every 10 minutes from 0:00:00 to 0:40:00
3. 360 of the Chow Hall at 0:17:00
4. Show of Force at Gate 2 starting at 0:20:00 and ending at 0:30:00
5. 15 minute Show of Force at Point Quebec
6. 360 of the Barracks at 0:53:00

RAMs for Scenario D

1. Show of Force at a point inside NAI 3 for 5 minutes at 0:44:00
2. Imagery of Gate 1 every 10 minutes from 0:10:00 to 0:50:00
3. 360 of the Chow Hall at 0:18:00
4. Show of Force at the Barracks starting at 0:15:00 and ending at 0:25:00
5. 15 minute LPOP at Point India
6. 360 of the Barracks at 0:52:00
**APPENDIX C. Sample Sensor Operator Script.**

<table>
<thead>
<tr>
<th>Minutes into Mission</th>
<th>Event</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>RAM 2: Imagery of Ammo Dump every 10 minutes from 0:00:00 to 0:40:00</td>
<td>1. Take image of ammo dump once the sensor feed is correctly positioned 2. Send the image to the C2 operator 3. Radio the C2 operator that the image has been sent and that no suspicious activity was observed</td>
</tr>
<tr>
<td>0</td>
<td>RAM 4: 360 of Barracks between 10:00 and 20:00</td>
<td>1. Take image of the barracks from the north, south, east, and west once the sensor feed is correctly positioned 2. Send the image to the C2 operator 3. Radio the C2 operator that the images have been sent and that no suspicious activity was observed</td>
</tr>
<tr>
<td>7</td>
<td>Unidentified Watercraft</td>
<td>1. Take an image of the boat once the sensor feed is correctly positioned 2. Send the image to the C2 operator 3. Radio the C2 operator that the image has been sent and that the watercraft is not a threat</td>
</tr>
<tr>
<td>~8</td>
<td>Sensor Failure</td>
<td>1. Radio C2 operator that vehicle X’s sensor feed is out</td>
</tr>
<tr>
<td>25</td>
<td>Overwatch Spartan-53 and Spartan-54</td>
<td>1. Take an image of the units once the sensor feed is correctly positioned 2. Send image to the C2 operator 3. Radio the C2 operator that the image has been sent and that the overwatch is in place</td>
</tr>
<tr>
<td>33</td>
<td>Gate Runner</td>
<td>1. Take an image of the flight line 2. Send image to the C2 operator 3. Radio the C2 operator that the image has been sent</td>
</tr>
<tr>
<td>50</td>
<td>Fence Alarm</td>
<td>1. Take an image of the fence once the UGV is in place 2. Send image to the C2 operator</td>
</tr>
<tr>
<td>55</td>
<td>Escort Orca-08</td>
<td>1. Take an image of Orca-08 once the sensor feed is correctly positioned 2. Send image to the C2 operator 3. Radio the C2 operator that the image has been sent and that the escort is in place</td>
</tr>
</tbody>
</table>
**APPENDIX D. Sample TOC Operator Script**

<table>
<thead>
<tr>
<th>Minutes into Mission</th>
<th>Event</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>~8</td>
<td>Sensor Failure</td>
<td>1. Kill the EO sensor of a vehicle that is on a play</td>
</tr>
<tr>
<td>16</td>
<td>Restore Sensor</td>
<td>1. Restore the EO sensor</td>
</tr>
<tr>
<td>20</td>
<td>ROZ</td>
<td>1. Create a ROZ that interferes with a UAV’s NCFP route (do not put ROZ over any blue force routes)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Send ROZ to C2 station</td>
</tr>
<tr>
<td>53</td>
<td>Flat Tire</td>
<td>1. Give the vehicle sent to inspect the fence alarm a flat tire</td>
</tr>
</tbody>
</table>
APPENDIX E. Background Questionnaire

Age_____ Gender_____

Experience with Unmanned Vehicle Operations (estimate):

<table>
<thead>
<tr>
<th>Vehicle Name</th>
<th>Non-Combat Flight Hrs.</th>
<th>Combat Flight Hrs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other Previous Flight Experience (Most recent first):

<table>
<thead>
<tr>
<th>Aircraft</th>
<th>Estimated number of flight hours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experience with Security Forces and/or Base Defense (estimate):

<table>
<thead>
<tr>
<th>Job/Deployment</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other related supervisory control or autonomous system development experience or skills?
APPENDIX F. Base Map
APPENDIX I. Overall System Questionnaire.

[SYSTEM] DEBRIEFING QUESTIONNAIRE

We would like to capture your general impressions of the overall [SYSTEM] system for supporting future scenarios involving multiple types of highly autonomous remotely piloted vehicles (UxVs). The [SYSTEM] concept includes all the display formats, symbology, control methods and ability to call plays verbally and manually, as well as tailor plan parameters.

1) Please indicate which response best matches your opinion of the POTENTIAL VALUE of the [SYSTEM] concept for future UxVs operations:

<table>
<thead>
<tr>
<th>Negative Impact</th>
<th>No Aid</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Great Aid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

2) Please indicate which response best matches your opinion of the ability of the [SYSTEM] concept to aid your WORKLOAD for future UxVs operations:

<table>
<thead>
<tr>
<th>Negative Impact</th>
<th>No Aid</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Great Aid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

3) Please indicate which response best matches your opinion of the ability of the [SYSTEM] concept to aid your SITUATION AWARENESS during future UxVs operations:

<table>
<thead>
<tr>
<th>Negative Impact</th>
<th>No Aid</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Great Aid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
**APPENDIX J. System Usability Scale (Brooke, 1996)**

**System Usability Scale**  
Page 1 of 2

Please answer the following questions by circling a single number from 1 (strongly disagree) to 5 (strongly agree).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I think that I would like to use this system frequently.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>2. I found the system unnecessarily complex.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>3. I thought the system was easy to use.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>4. I think that I would need the support of a technical person to be able to use this system.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>5. I found the various functions in this system were well integrated.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>6. I thought there was too much inconsistency in this system.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>7. I would imagine that most people would learn to use this system very quickly.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>8. I found the system very cumbersome to use.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>9. I felt very confident using the system.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>10. I needed to learn a lot of things before I could get going with this system.</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
</tbody>
</table>

Please answer the following questions with your honest opinion, providing as much detail as possible:

11. What did you like most about the system?
12. What did you like least about the system?

13. What was confusing about the system?

14. How would you improve the system?

15. What was inconsistent about the system?
APPENDIX K. IMPACT vs. Baseline Questionnaire

Performance on each task (different task in each row of table): Next, compare your performance on each task between the two conditions (e.g., how often did you think you correctly answered commander queries with and without IMPACT?).

<table>
<thead>
<tr>
<th>Task</th>
<th>Performance Better with IMPACT</th>
<th>Performance Slightly Better with IMPACT</th>
<th>Performance Equal (similar to)</th>
<th>Performance Slightly Better With BASELINE</th>
<th>Performance Better With BASELINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Full Coverage Patrol</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Going Highly Mobile</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Random Anti-terror Measures</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Avoid Restricted Operating Zones</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Get Eyes-On Suspicious Vehicle</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Provide Overwatch</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Perform Scout-Ahead</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Perform Cordon</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Perform Blockade</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Respond to Intruder Events</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Respond to Commander Queries</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Respond to Vehicle Failures</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Respond to Environmental Events</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments: