



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**STRATEGY TO IMPROVE THE TRUST BETWEEN HUMANS
AND ARTIFICIAL INTELLIGENCE ENABLED AIR AND
MISSILE DEFENSE (AMD) SYSTEMS**

by

Ming Hui Peh

September 2021

Thesis Advisor:
Second Readers:

Bonnie W. Johnson
Walter A. Kendall
John M. Green

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 2021		3. REPORT TYPE AND DATES COVERED Master's thesis
4. TITLE AND SUBTITLE STRATEGY TO IMPROVE THE TRUST BETWEEN HUMANS AND ARTIFICIAL INTELLIGENCE ENABLED AIR AND MISSILE DEFENSE (AMD) SYSTEMS			5. FUNDING NUMBERS	
6. AUTHOR(S) Ming Hui Peh				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) Artificial intelligence (AI) has the potential to alter many aspects of military operations and improve overall operational effectiveness. One particularly complex mission domain for the U.S. Department of Defense (DOD) is air and missile defense (AMD). With the proliferation of more advanced weapons, there is a greater need for warfighters to quickly assess the situation, develop an appropriate course of action, and best utilize their warfare assets to respond. This series of activities requires warfighters to have a high level of trust in the system. However, trust in AI systems is not universally defined, and there is no common criteria for evaluating an AI system's trustworthiness. This thesis studies how the established trust factors in the literature could apply to the AMD domain to enable and enhance trust between human operators and future AI-AMD systems, and how trustworthiness can be designed into future AI-AMD systems. The thesis proposes a framework of trust and human-machine interactions (HMI) in AI-AMD systems. Thereafter, the thesis proposes a set of trust factors considering the operational and organization environment, as well as the operator and AI-AMD decision-aid team dynamics. This thesis uses the U.S. military solutioning framework of DOTMLPF-P (Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, Facilities, and Policy) to develop a strategy to improve calibrated trust between the operator and AI-AMD system.				
14. SUBJECT TERMS artificial intelligence, air and missile defense system, trust			15. NUMBER OF PAGES 181	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**STRATEGY TO IMPROVE THE TRUST BETWEEN HUMANS
AND ARTIFICIAL INTELLIGENCE ENABLED AIR AND MISSILE
DEFENSE (AMD) SYSTEMS**

Ming Hui Peh
Captain, Singapore Army
B Ch E, National University of Singapore, 2015

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING

from the

**NAVAL POSTGRADUATE SCHOOL
September 2021**

Approved by: Bonnie W. Johnson
Advisor

Walter A. Kendall
Second Reader

John M. Green
Second Reader

Oleg A. Yakimenko
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Artificial intelligence (AI) has the potential to alter many aspects of military operations and improve overall operational effectiveness. One particularly complex mission domain for the U.S. Department of Defense (DOD) is air and missile defense (AMD). With the proliferation of more advanced weapons, there is a greater need for warfighters to quickly assess the situation, develop an appropriate course of action, and best utilize their warfare assets to respond. This series of activities requires warfighters to have a high level of trust in the system. However, trust in AI systems is not universally defined, and there is no common criteria for evaluating an AI system's trustworthiness.

This thesis studies how the established trust factors in the literature could apply to the AMD domain to enable and enhance trust between human operators and future AI-AMD systems, and how trustworthiness can be designed into future AI-AMD systems. The thesis proposes a framework of trust and human-machine interactions (HMI) in AI-AMD systems. Thereafter, the thesis proposes a set of trust factors considering the operational and organization environment, as well as the operator and AI-AMD decision-aid team dynamics. This thesis uses the U.S. military solutioning framework of DOTMLPF-P (Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, Facilities, and Policy) to develop a strategy to improve calibrated trust between the operator and AI-AMD system.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION	1
A.	STATEMENT OF NEED.....	1
	1. Problem Statement.....	2
	2. Thesis Objectives.....	2
B.	BACKGROUND AND MOTIVATION	3
C.	BENEFITS OF THE STUDY	4
D.	RESEARCH METHOD AND THESIS ORGANIZATION.....	5
II.	LITERATURE REVIEW – TRUST IN AI SYSTEMS	7
A.	DEFINING AI SYSTEM.....	7
B.	DEFINING TRUST IN AI SYSTEM.....	8
	1. Default Trust, Impersonal Trust and Human-AI Trust.....	8
	2. Contractual Trust	10
	3. Trust Repair	10
	4. Trustworthy AI	11
	5. Trust Calibration, Resolution and Specificity.....	11
C.	HUMAN-MACHINE INTERACTIONS IN ARTIFICIAL INTELLIGENCE SYSTEMS.....	13
	1. Characterization of Human-Machine Interactions	14
	2. Levels of Machine Automation and Human Information Processing – Deep Dive into “Selecting an Action” Stage of Human-Machine Interactions	16
	3. Human-Machine Interactions Design Constructs.....	21
	4. Joint Activity Theory of Human-Machine Interactions.....	24
	5. Automation Taxonomy for HMI in Air Traffic Management	26
D.	TRUST FACTORS IN AI SYSTEMS.....	31
	1. Schaefer et al.’s (2016) Three-Factor Trust Model.....	31
	2. Objective/Subjective Factors of Trust and Trustworthiness.....	44
	3. Denning’s Six-User level (subjective) Software Quality Assessment	46
E.	TRUST REPAIR FACTORS IN AI SYSTEMS	48
	1. Relationship Act.....	48
	2. Relationship Regulation Act	49
	3. Net Victim Effect.....	49
	4. Repair Strategies	49
F.	KEY TAKEAWAYS FROM CHAPTER II.....	51

III.	LITERATURE REVIEW – OVERVIEW OF AI SYSTEMS FOR AIR AND MISSILE DEFENSE.....	55
A.	OVERVIEW OF AI-AMD THREAT SCENARIOS.....	55
1.	Specific Threats.....	55
2.	AMD Threat Scenarios.....	56
B.	OVERVIEW OF AI-AMD OPERATIONS AND ARCHITECTURE.....	59
1.	Operational Viewpoints and Scenarios.....	59
2.	Characteristics of AI-AMD System.....	74
C.	REVIEW OF U.S. ARMY PATRIOT MISSILE SYSTEM	75
D.	KEY TAKEAWAYS FROM CHAPTER III	79
IV.	ANALYSIS OF TRUST, HUMAN-MACHINE INTERACTIONS (HMI) AND TRUST FACTORS OF AI-AMD SYSTEM.....	81
A.	SYSTEM BOUNDARIES AND ENVIRONMENT	81
1.	AI-AMD C2 System	82
2.	Sensors	84
3.	BLUFOR Weapon System	84
4.	Operational Environment	84
5.	Organizational Environment	85
6.	Threats	85
B.	TRUST CONCEPTUAL FRAMEWORK AND DEFINITIONS	86
1.	Trust.....	87
2.	AI-AMD Decision Aid Capability and Trustworthiness	87
3.	Default Trust	88
4.	Optimum Level of Trust/Calibrated Trust, Trust Violations and Repair	88
C.	HUMAN-MACHINE INTERACTIONS BETWEEN HUMAN OPERATORS AND AI-AMD DECISION AIDS	89
1.	Functional Allocations between Operator and AI-AMD Decision Aid	89
2.	Proposed Level of Automation (LOA) and Functional Allocation	95
3.	Operator and AI-AMD Decision Aid HMI.....	100
4.	Design Constructs of Operator and AI-AMD Decision Aid HMI.....	101
5.	Implications of the HMI on Trust	103
D.	TRUST FACTORS OF HUMANS AND AI-AMD DECISION AID	104
1.	Operational Environment and Context	105
2.	Organizational Environment	107

3.	AI-AMD C2 System: Operator-Related Trust Factors	108
4.	AI-AMD C2 System: AI-AMD Decision Aid Trustworthiness.....	110
5.	“Team” Related Factors.....	112
V.	PROPOSED STRATEGY TO ACHIEVE CALIBRATED TRUST BETWEEN THE OPERATOR AND AI-AMD DECISION AID	117
A.	SOLUTIONS TO ACHIEVE CALIBRATED TRUST BETWEEN OPERATOR AND AI-AMD DECISION AID.....	117
1.	Doctrine – Shift in Organizational Paradigm	118
2.	Increase in Training Requirements.....	120
3.	Materiel – AI-AMD Decision Aid System Development	124
4.	Organization	130
5.	Leadership and Education	131
6.	Personnel.....	131
7.	Facilities	132
8.	Policy	132
B.	SOLUTIONS TO REPAIR TRUST	133
VI.	CONCLUSION AND FUTURE RESEARCH.....	135
A.	KEY INSIGHTS.....	135
B.	FUTURE WORK.....	137
	LIST OF REFERENCES	139
	INITIAL DISTRIBUTION LIST	153

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

Figure 1.	Thesis Methodology.....	6
Figure 2.	Schematic of the Flow of Chapter II.....	7
Figure 3.	Framework of Trust. Adapted from Ajzen and Fishbein (1977).	9
Figure 4.	Relationship between Trust Calibration and Automation Capability. Source: Lee and See (2004).	12
Figure 5.	Stages, Design Considerations and Support Tools of HMI Activities. Adapted from Norman (1984).	15
Figure 6.	Four-Stage Model of Human Information Processing. Adapted from Parasuraman, Sheridan, and Wickens (2000).	18
Figure 7.	Mapping of System Functions to Four-Stage Model of Human Information Processing. Source: Parasuraman, Sheridan, and Wickens (2000).	19
Figure 8.	Flow Chart to Guide Levels of Automation Design. Source: Parasuraman, Sheridan, and Wickens (2000).	20
Figure 9.	Five Constructs of HMI. Adapted from Pritchett and Feary (2011).	21
Figure 10.	Three-factor Trust Model. Adapted from Schaefer et al. (2016).	32
Figure 11.	Trust Repair Model. Source: de Visser, Pak, and Shaw (2018).	48
Figure 12.	Schematic of the Flow of Chapter III.	55
Figure 13.	OV-1 of AMD Operations	60
Figure 14.	OV-6c: AI-AMD Event Trace for AMD Operations. Source: Jones et al. (2020).	62
Figure 15.	OV-5a: AI-AMD Operational Activity Decomposition - Internal Capabilities. Source: Jones et al. (2020).	64
Figure 16.	OV-5a: AI-AMD Operational Activity Decomposition – External Systems. Source: Jones et al. (2020).	65
Figure 17.	OV-5b: AI-AMD Operational Model. Source: Jones et al. (2020).	66
Figure 18.	OV-5b: Decomposed AI-AMD Operational Activity Model. Source: Jones et al. (2020).	67

Figure 19.	OV-5b/6c: AI-AMD Action Diagram. Source: Jones et al. (2020).	68
Figure 20.	Decomposition of F2T2EA Find Step. Source: Jones et al. (2020).	70
Figure 21.	Decomposition of F2T2EA Fix Step. Source: Jones et al. (2020).	71
Figure 22.	Decomposition of F2T2EA Track Step. Source: Jones et al. (2020).	71
Figure 23.	Decomposition of F2T2EA Target Step. Source: Jones et al. (2020).	72
Figure 24.	Decomposition of F2T2EA Engage Step. Source: Jones et al. (2020).	73
Figure 25.	Decomposition of F2T2EA Engage Step. Source: Jones et al. (2020).	73
Figure 26.	Patriot Missile System Decision Cycle. Source: Hawley (2017).	76
Figure 27.	AI-AMD System Context Diagram.	82
Figure 28.	Trust between Human Operator and AI-AMD Decision Aid.	86
Figure 29.	NASA-TLX Rating Scale and Definition. Source: Hart (2006).	93
Figure 30.	Schematic of the HMI of “Team” Functions. Source: Hawkey, Mares, and Giammanco (2005).	100
Figure 31.	Summary of Trust Factors	105
Figure 32.	Strategy Articulation Map to Achieve Calibrated Trust between Operator and AI-AMD Decision Aid	118
Figure 33.	Proposed Training Regime for AI-AMD Operators	121

LIST OF TABLES

Table 1.	10-point Scale for Automation of Decision and Action Selection. Source: Parasuraman, Sheridan, and Wickens (2000).	16
Table 2.	HMI to Enhance Performance. Adapted from Bradshaw, Feltovich, and Johnson (2012).	25
Table 3.	Eight Maxims for Good Human or Machine Agent. Source: Bradshaw, Feltovich, and Johnson (2012).	26
Table 4.	Proposed New Taxonomy for Air Traffic Management (Part I). Source: Save, Feuerberg, and Avia (2012).	28
Table 5.	Proposed New Taxonomy for Air Traffic Management (Part II). Source: Save, Feuerberg, and Avia (2012).	29
Table 6.	Proposed New Taxonomy for Air Traffic Management (Part III). Source: Save, Feuerberg, and Avia (2012).	30
Table 7.	Objective Factors of Trust/Trustworthiness. Adapted from Sheridan (1988).	44
Table 8.	Subjective Factors of Trust/Trustworthiness. Adapted from Sheridan (2019).	45
Table 9.	Denning’s Six-User Level Software Quality Assessment. Adapted from Denning (2016).	47
Table 10.	Trust Repair Types and Descriptions. Source: De-Visser, Pak, and Shaw (2018).	50
Table 11.	Summary of AMD Threats. Adapted from Jones et al. (2020), Airforce-technology (2011) and Airforce-technology (2021).	56
Table 12.	Mapping of Critical Functions to F2T2EA and OODA Loop	63
Table 13.	Risk Assessment Criteria. Adapted from DOD (2017).	91
Table 14.	Summary of Kill Chain Functions and Decision Risk Values. Adapted from Jones et al. (2020).	91
Table 15.	NASA-TLX Rating across 17 Kill Chain Functions.	94
Table 16.	Cognitive Workload Criteria.....	94
Table 17.	Summary of Kill Chain Function Allocations.	96

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
AI-AMD	Artificial Intelligence-enabled Air and Missile Defense
AMD	Air and Missile Defense
BDA	Battle Damage Assessment
BLUFOR	Blue Force
COA	Course-of-Action
CONOPs	Concept of Operations
C2	Command and Control
DARPA	Defense Advanced Research Projects Agency
DOD	Department of Defense
DoDAF	Department of Defense Architecture Framework
DOTMLPF-P	Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, Facilities and Policy
F2T2EA	Find-Fix-Track-Target-Engage-Assess
ICOM	Inputs-controls-outputs-mechanisms
JP	Joint Publication
LOA	Level of Automation
OODA	Observe-Orient-Decide-Act
HMI	Human-Machine Interactions
NASA-TLX	National Aeronautics and Space Administration-Task Load Index
RAM	Rockets, artillery and mortars
REDFOR	Red Force
SA	Situation Awareness
SoS	System-of-Systems
THAAD	Terminal High Altitude Area Defense
TTPs	Training, Tactics and Procedures
UAV	Unmanned Aerial vehicle
VUCA	Volatile, Uncertain, Complex and Ambiguous
XAI	Explainable Artificial Intelligence

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Currently, artificial intelligence (AI) offers tremendous opportunities to transform many domains of military operations, including warfighting, command and control (C2), logistics, security, and maintenance, to enhance their overall operational effectiveness. One particularly complex mission domain where the application of AI is critical is air and missile defense (AMD). The AMD mission refers to the defense of the homeland, protected areas, ground bases, ground forces, or ships against an adversarial air or missile threat. AMD threats include hostile aircraft, unmanned aerial vehicles (UAVs), or airborne missiles. Complexity in AMD operations stems from threat severity, the unexpected nature of the threat, uncertainty in situational knowledge, and the fast tempo of events as warfighters must quickly assess the situation, develop an appropriate course of action, and best utilize their warfare assets to respond. The United States Department of Defense (U.S. DOD) is studying the use of AI systems (or AI-enabled AMD [AI-AMD] systems) as automated decision aids for AMD warfighters to significantly reduce their cognitive load (Jones et al. 2020), enabling faster and better AMD decisions.

One critical aspect of AI which has gathered a lot of research interest is trust. Trust is an essential tenet for effective teams. It applies to both humans and human-machine teams. Trust allows team members to work together well and is the foundation for effective team performance (Lee and See 2004). Successful collaboration with AI systems will necessitate humans having a calibrated level of trust and reliance on AI systems (Taylor et al. 2016).

With the development and fielding of more advanced and faster air and missile threat munitions, there is a greater need for operators to make decisions quickly during AMD operations. Untimely decisions and responses will result in catastrophic consequences. Therefore, AI is a possible solution to speed up and enhance the decision-making process through automated decision aids. These AMD automated battle management aids can help tactical operators cope with quicker decision cycles, massive volumes of data, and several systems or screens to watch (Galdorisi 2019). However, to effectively leverage the potential of AI capabilities, a high degree of operator trust is

needed. A low level of operator trust in the system may cause the AI-AMD system to be underutilized, unduly monitored, or not used at all (Floyd, Drinkwater, and Aha 2016). Any of these problems could result in an unnecessary increase in the operator's workload or the possibility of mission failure.

The thesis carried out an extensive review on the definitions of trust, concepts of human-machine interactions (HMI), trust factors, and conceptual models including models of AMD kill chains, threat scenarios, architectures, models, and functions. With this understanding, the thesis proposed a trust framework for the AI-AMD system, description of HMI and AI-AMD system trust factors. The thesis concluded with a strategy to achieve calibrated trust between the human operators and AI-AMD systems.

The trust framework began with an analysis of the system context. Figure 1 shows a context diagram of the AI-AMD Command and Control (C2) system (which comprises of AI-AMD operator and decision aid) and other sub-systems it interacts with, that contribute to the development of trust between the operator and the AI-AMD decision aid. The context diagram allowed us to study the interactions amongst systems and their influences on the trust dynamics between the AI-AMD operators and decision aid.

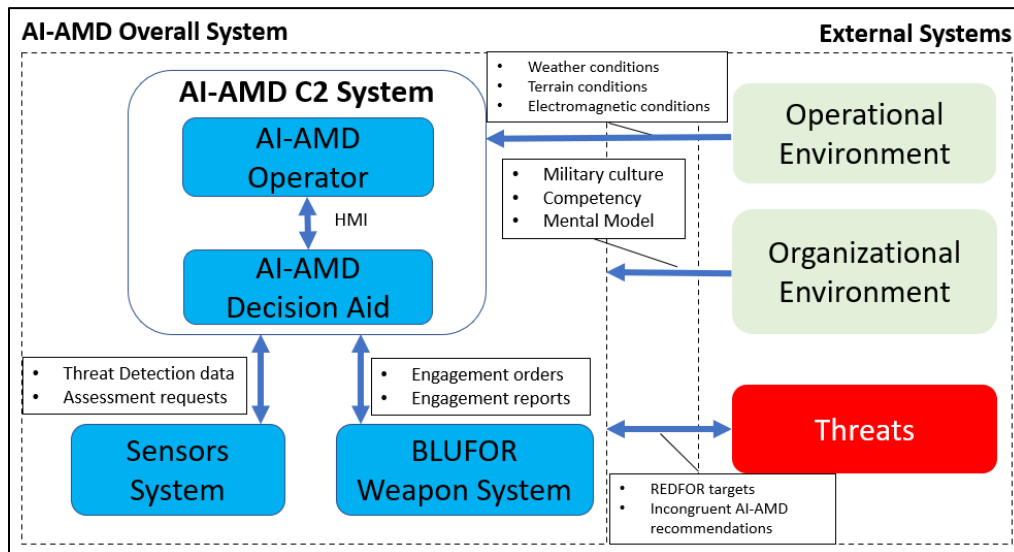


Figure 1. AI-AMD System Context Diagram.

The thesis defined trust as the operator's attitude that an AI-AMD decision aid will help to achieve the operator's goals of accomplishing the mission of destroying incoming threats promptly in an operational environment that is characterized by uncertainty and vulnerability. This definition of trust indicates that it is an emotional assessment of what the operator believes or perceives of the AI-AMD decision aid. To positively influence trust, the operator must personally see and feel the advantages of the AI-AMD decision aid actions. The AI-AMD operations involve a lot of uncertainties and vulnerabilities arising from the environmental factors of weather, electromagnetic interferences, and terrain as well as the nature of the ever-evolving threats. The operator will anticipate that the AI-AMD decision aid performs as per "contracts" to deal with these uncertainties and vulnerabilities. The contracts will be the perceived functions or tasks that the AI-AMD decision aid is supposed to carry out as well as the desirable performance of carrying out these functions or tasks.

Figure 2 illustrates the trust framework between the operators and AI-AMD decision aid. The y-axis represents the level of trust human has for the AI-AMD decision aid while the x-axis represents the capability of the AI-AMD decision aid. The green dotted 45° line shows the optimum trust level or calibrated trust line, where trust corresponds to AI-AMD capabilities, resulting in proper utilization (Lee and See 2004). Over-trust is represented by the region above the optimum trust level line, when trust exceeds system capabilities, resulting in misuse. The zone below the ideal level of trust shows distrust, where trust falls short of system capabilities, resulting in disuse. It was assumed that there exists some level of default trust (shown as the yellow box), the goal is to develop a strategy to improve the trust given an AI-AMD decision aid capability. Trust violations might occur that reduce the operator's trust when operating with the system. Hence, trust repair actions must be incorporated to maintain the trust at an optimum level.

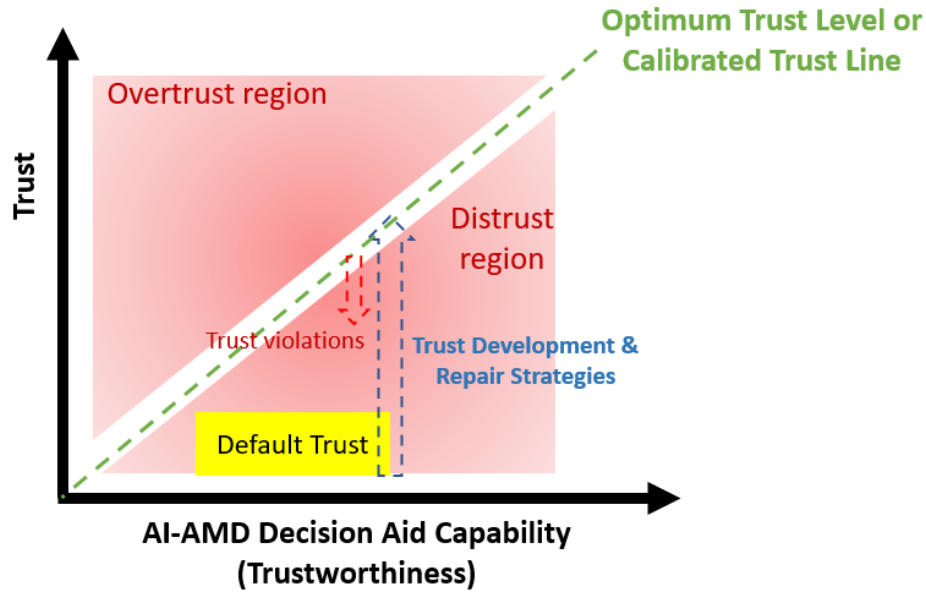


Figure 2. Trust Between Human Operator and AI-AMD Decision Aid.

The study on the HMI between the operator and AI-AMD decision aid based on functional allocations suggested that operators and AI-AMD decision aid should be designed to operate as a “team” across most of the kill chain functions. This directed the thesis to study trust factors that consider both the humans and decision aid dimensions individually and collectively. The study on the HMI between the operators and AI-AMD decision also revealed that the operators’ role has changed from manual controller to supervisory controller. A trustworthy decision aid is thus important as the operator would expect that the system performs as expected to help the operator perform his role better. Also, to further help ease the cognitive workload of the operator, the external representation of the information, recommendations from the decision aid must be easily readable and understood.

On trust factors, the thesis proposed an “outside-in” framework as shown in Figure 3. The thesis first considered factors associated with the operational environment which described the context of the AMD operation system. Second, it studied factors associated with the organizational environment where the operators and AI-AMD systems have been trained and developed. Third, the thesis studied the interactions between the operators and

AI-AMD decision aids to propose factors associated with both the operators, the AI-AMD, individually and collectively.

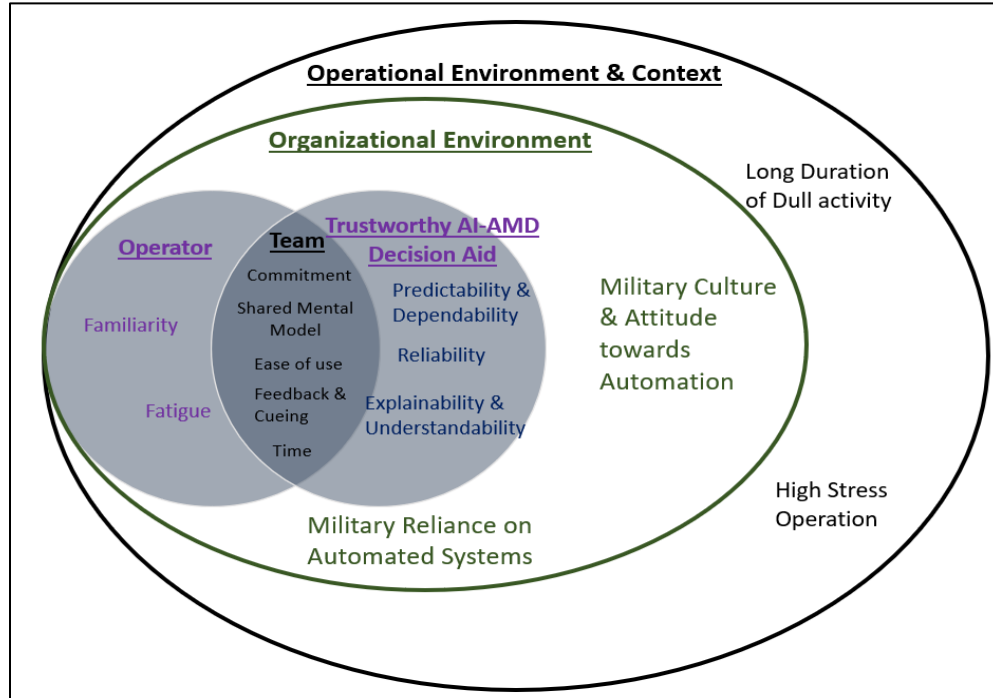


Figure 3. Proposed Trust Factors

Figure 4 shows the proposed strategy articulation map to achieve calibrated trust between operator and AI-AMD decision aid. The review of the trust definitions, HMI, and trust factors suggested that the strategy should focus on three key areas: (1) collective and individual humans' perception of automation technology and AI-AMD decision aid, (2) enhancing operator and AI-AMD decision aid team dynamics, and (3) trustworthiness of AI-AMD decision aid focusing on system development. The strategy proposed three key tenets and five supporting tenets using the DOTMLPF-P framework. First, military warfighters need to be informed of the true capabilities and limitations of automation technology, particularly AI-AMD decision aid. Second, the operator's training requirements must increase to cope with a new job scope and evolving threats. Third, new requirements must be added to the AI-AMD decision aid system development to improve the system's perceived trustworthiness. These three key tenets are supported by the other

aspects of the DOTMLPF-P framework such as organization, leadership, personnel, facilities, and policy.

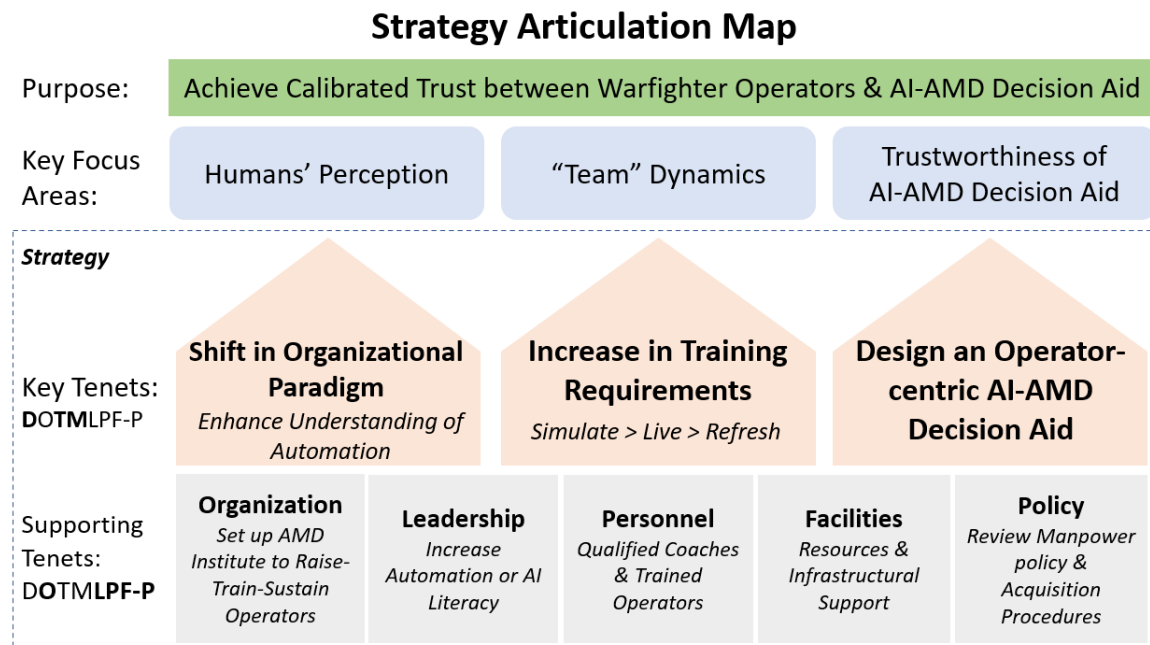


Figure 4. Strategy Articulation Map to Achieve Calibrated Trust Between Operator and AI-AMD Decision Aid

References

- Floyd, Michael W, Michael Drinkwater, and David W Aha. 2016. "Learning Trustworthy Behaviors using an Inverse Trust Metric." In *Robust Intelligence and Trust in Autonomous Systems*, 33–53. Springer.
- Galdorisi, George. 2019. "The Navy Needs AI, It's Just Not Certain Why." USNI Proceedings.
- Jones, J. Issac, Russell Kress, Adam Rahman, and William Newmeyer. 2020. "Leveraging Artificial Intelligence (AI) for Air and Missile Defense (AMD): An Outcome Oriented Decision Aid." Master's thesis, Naval Postgraduate School. <https://calhoun.nps.edu/handle/10945/66088>.

Lee, John D, and Katrina A See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80.

Taylor, Gavin, Ranjeev Mittu, Ciara Sibley, and Joseph Coyne. 2016. "Towards Modeling the Behavior of Autonomous Systems and Humans for Trusted Operations." In *Robust Intelligence and Trust in Autonomous Systems*, 11–31. Springer.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to first express my heartfelt thanks to Professor Bonnie Johnson, my thesis advisor, and Professor John Mike Green and Professor Walter Anthony Kendall, my thesis co-advisors, for their valuable input and advice provided throughout this journey. Their support and guidance have been instrumental in improving the content and flow of my thesis, and this effort would not have been possible without the strong support and guidance from them.

I would like to express my gratitude to my wife and children, Rowenna, Owen, and Gwen, as well as my extended family and loved ones, for their support and encouragement throughout this successful journey. Thank you for the support in my endeavor to further my studies overseas and achieve a milestone in my academic aspirations.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. STATEMENT OF NEED

Machines have been a great and effective team player for the United States military. Over the years, the machines adopted by the U.S. military have evolved from simple remote-controlled vehicles to complicated autonomous artificial intelligence (AI) systems such as unmanned aerial systems (UAS) to support its missions in ever-increasingly volatile, uncertain, complex, and ambiguous (VUCA) operational environments. Recently, the U.S. Department of Defense (DOD) has recognized the need for the swift adoption of AI to undertake a greater number of mission areas to improve future operational capabilities (Galdorisi 2019). The DOD has established clear direction with several roadmaps to enhance the levels of autonomy applied in the U.S. military (Winnefield and Kendall 2012). Machines are no longer just tools; they are becoming teammates with increasing intelligence and autonomy.

One particularly complex mission domain for the U.S. DOD is air and missile defense (AMD). The AMD mission refers to the defense of the homeland, protected areas, ground bases, ground forces, or ships against an adversarial air or missile threat. AMD threats include hostile aircraft, unmanned aerial vehicles (UAVs), or airborne missiles. Complexity in AMD operations stems from threat severity, the unexpected nature of the threat, uncertainty in situational knowledge, and the fast tempo of events as warfighters must quickly assess the situation, develop an appropriate course of action, and best utilize their warfare assets (sensors, communications, weapons, and countermeasures) to respond. The U.S. DOD is studying the use of AI systems (or AI-enabled AMD [AI-AMD] systems) as automated decision aids for AMD warfighters to significantly reduce their cognitive load (Jones et al. 2020), enabling faster and better AMD decisions.

Trust is an essential tenet for effective teams. It applies to both humans and human-machine teams. Trust allows team members to work together well, and hence, is the foundation for effective team performance (Lee and See 2004). However, trust in AI systems is not unanimously well understood or well defined (Toreini et al. 2020). There is

no standard metric that can be used to describe trust in AI systems and to determine the trustworthiness of an AI system. Successful collaboration with AI systems will necessitate humans having a calibrated level of trust and reliance on AI systems (Taylor et al. 2016).

1. Problem Statement

With the development and fielding of more advanced and faster air and missile threat munitions, there is a greater need for operators to make decisions quickly during AMD operations. Untimely decisions and responses will result in catastrophic consequences. Therefore, AI is a possible solution to speed up and enhance the decision-making process through automated decision aids. These AMD automated battle management aids can help tactical operators cope with quicker decision cycles, massive volumes of data, and several systems or screens to watch (Galdorisi 2019). However, to effectively leverage the potential AI capabilities, a high degree of operator trust is needed. A low level of operator trust in the system may cause the AI-AMD system to be underutilized, unduly monitored, or not used at all (Floyd, Drinkwater, and Aha 2016). Any of these problems could result in an unnecessary increase in the operator's workload or the possibility of mission failure.

2. Thesis Objectives

This thesis studied how the established trust factors in the literature can apply to the AMD domain to enable and enhance trust between human operators and future AI-AMD systems; and how trustworthiness can be engineered into future AI-AMD systems.

The thesis objectives are as follows:

- Understand and describe trust and human-machine interactions (HMI) of future AI-AMD systems.
- Conduct an in-depth study of trust factors and their impacts on the kill chain functions of the conceptual AI-AMD system.
- Develop AMD threat scenarios to study the operational outcomes of variances in trust.
- Propose a strategy to improve trust between the operator and AI-AMD system.

B. BACKGROUND AND MOTIVATION

(1) Introduction of AI to Military Operations

Over the past decade, AI has gained tremendous attention (Palmer, Selwyn, and Zwillinger 2016). AI has been adopted in many government services, businesses, and industries. As AI technology matures, it is increasingly poised to be applied to military operations (Scharre 2018). Currently, AI offers tremendous opportunities to transform many domains of military operations, including warfighting, command and control (C2), logistics, security, and maintenance, to enhance their overall operational effectiveness. As such, the U.S. DOD has recognized the need to adopt AI technology and has developed an AI Strategy (U.S. DOD 2018) to hasten the adoption of AI across military missions and domains.

(2) Adoption of Five AI Ethical Principles by DOD

While there are huge benefits from leveraging AI, its adoption has also initiated numerous debates on ethical and trust aspects due to its infancy, ambiguity, and risks. The U.S. DOD has acknowledged these challenges and in February 2020, officially announced its development of DOD ethical principles, namely that AI systems be: (1) equitable, (2) responsible, (3) reliable, (4) traceable, and (5) governable to guide the ethical and lawful adoption of AI (U.S. DOD 2020). This news indicates that the DOD is serious about *how* AI capabilities are used, and not just setting a course for the unrestrained adoption of AI (Lamberth 2020).

(3) Poor Understanding of Trust in AI Systems

However, there is a challenge in translating the five ethical principles into frameworks that can be institutionalized across the various services and missions within the DOD that guide the engineering of operational systems (Lamberth 2020). For instance, trust in AI systems is a critical tenet that is linked to the five principles. Trust in AI systems is not unanimously well understood nor well defined (Toreini et al. 2020). No standard metric exists to evaluate the trustworthiness of an AI system or how the level of trustworthiness may affect system performance.

(4) Trust in AI-AMD Systems

The AMD mission is critical to preserving and protecting defended areas and DOD forces against adversaries' missiles, aircraft, and UAVs. With the fielding of more advanced weapons such as loitering munitions and hypersonic missiles, AMD operations become more complex and time-sensitive in the ability to quickly identify, assess, and neutralize these threats. There is a need for warfighters to make precise and fast decisions in response to these threats. A recent NPS project (Jones et al. 2020) analyzed the usage of AI to enhance situation awareness (SA) and battle management decision-making in the time-sensitive AMD domain. It was concluded that the application of AI will reduce the time required to respond to a threat significantly, expediting the AMD kill chain. However, the tenets of trust to determine the trustworthiness of an AI-AMD system that relates to the ethical principles were not considered. There is a need to understand this to develop a strategy to engineer a trustworthy system that meets both its operational needs and adheres to ethical principles.

(5) Importance of Trust – Remembering the USS Vincennes Incident

On July 3, 1988, the USS Vincennes accidentally shot down Iran Air's Commercial Flight 655. During this incident, the USS Vincennes' radar had identified a plane, and a follow-on standard Identification Friend or Foe (IFF) query identified the plane as a commercial airliner. The commanding officer, Capt. Will C. Rogers III relied on his own judgement and deduced that the plane was an F-14 and was diving towards USS Vincennes (CG-49). A nearby ship, the USS Sides (FFG-14), which was covering the same airspace, received the same radar and IFF information concerning the Iranian plane from its systems; yet it did not react in the same manner. The commanding officer of USS Sides acted differently as he responded primarily to the data provided on the radar screen (McCabe 2014). This incident highlighted the complex dynamics of humans-machines trust and the consequences of wrong perceptions of the data.

C. BENEFITS OF THE STUDY

This thesis is a continuation of the overarching direction that the DOD is undertaking in the development and institutionalization of AI ethical principles across all

services. Specifically, this thesis will increase the understanding of the trust factors in an AI-AMD system. The proposed trust strategy will provide us insights into developing trustworthy AI-AMD systems that meet operational requirements while not violating any ethical principles. This thesis is also an extension of the DOD's goal of implementing aligned AI ethical principles into AI systems across its many mission domains. The trust strategy will also provide insights that can be applied to other non-military AI systems.

D. RESEARCH METHOD AND THESIS ORGANIZATION

(1) Research and Analysis Method

This thesis applied an analysis and design research method, as illustrated in Figure 1. The study began with a literature review on AI and trust that explored researchers' work on human-machine interactions, trust factors, and methods for engineering trust that are applicable to existing AI systems. Next, the literature review explored AMD, including models of AMD kill chains, threat scenarios, architectures, models, and functions in relation to the human-machine interactions and trust factors that have been developed for other AI systems.

In the second stage of research, the study applied a systems analysis approach to identify human-machine interactions (HMI) in each of the kill chain functions and analyze each in terms of the identified trust factors such as complexity of the situation, predictability, explainability, reliability, and decision risks, etc.

In the third and final phase of research, the study identified and analyzed solutions to form a strategy to achieve calibrated trust between human and AI-AMD system and the development of a trustworthy AI-AMD system.

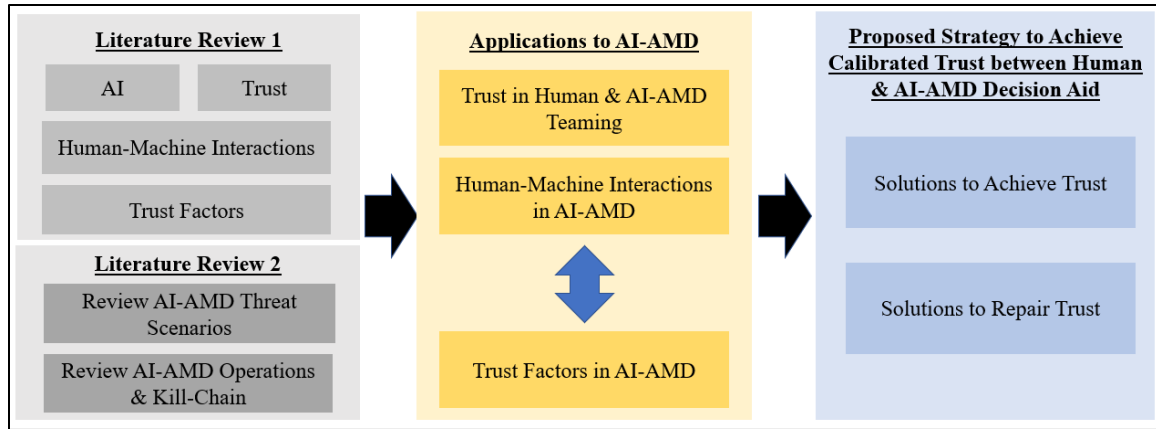


Figure 1. Thesis Methodology

(2) Thesis Organization

This thesis is divided into six chapters, including the current Chapter I: Introduction. Chapter II consists of a comprehensive literature review of the existing work done on trust, human-machine interactions and trust models and factors on existing systems. Chapter III contains an overview of the AI-AMD threat scenarios and operations and its conceptual system architecture. Chapter IV describes the trust in human and AI-AMD teaming, human-machine interactions that exist in the AI-AMD operations. It also lays out the trust factors that are applicable to the AI-AMD operations. Chapter V discusses the strategies to improve trust between the operator and AI-AMD system. Finally, Chapter VI concludes the collaborative study with comments and recommendations for future research in trust modeling of human-machine interactions and AI-AMD.

II. LITERATURE REVIEW – TRUST IN AI SYSTEMS

This chapter presents the first part of the extensive literature review which explored trust and its relationship with AI systems. Figure 2 shows the flow of Chapter II. First, AI was defined to better understand its characteristics and capabilities. After understanding what an AI system is, an in-depth literature review on the human-machine interaction (HMI) was carried out where the “machine” in this thesis is regarded as an AI system. The insights gained from this review form the basis for later describing the HMI of an AI-AMD system.

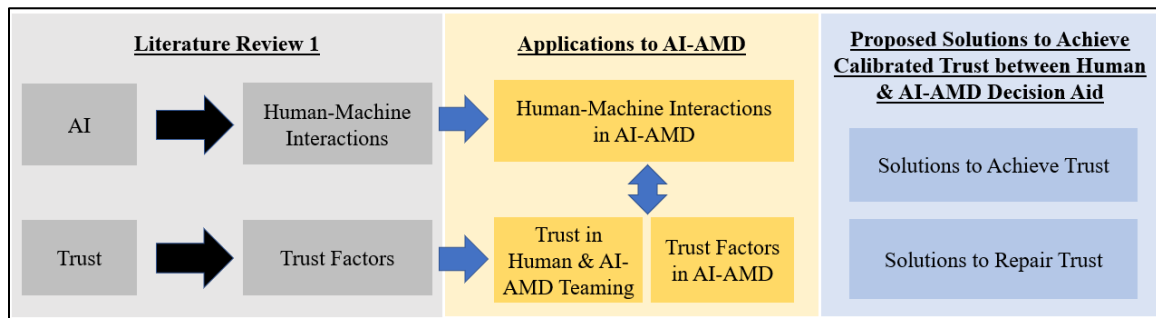


Figure 2. Schematic of the Flow of Chapter II.

Next, the various definitions in the literature of trust and its related trust elements were reviewed to deepen the understanding of human-AI trust. The chapter concludes with an in-depth review of the possible factors that affect trust between humans and AI systems. The insights gained from this literature review provide a foundation of knowledge for understanding trust between operators and AI-AMD system, and for postulating possible human-AI trust factors that are important in an AI-AMD system. Establishing trust factors for future AI-AMD systems is key to engineering trustworthy systems and achieving effective HMI for trust development.

A. DEFINING AI SYSTEM

This thesis adopted the DOD’s definition of AI: “the ability of machines to perform tasks that normally require human intelligence—for instance, recognizing patterns,

learning from experience, drawing conclusions, making predictions, or taking action—whether digitally or as the smart software behind autonomous physical systems” (U.S. DOD 2018, 5). From the DOD’s perspective, an AI system is any automation system that possesses a smart computer agent who can perform tasks with human-like intelligence that is attributed with the intent of a user. This type of intelligence is however different from biological human intelligence as it is built with machine intelligence. The AI-AMD system studied into this thesis fits in to this definition as the AI developed by humans supports the tasks listed by the DOD from recognizing pattern to making predictions in aiding the operators to act more swiftly.

B. DEFINING TRUST IN AI SYSTEM

This section reviews various established definitions of impersonal trust and human-automation trust to define human-machine trust with the AI-AMD domain. Subsequently, different elements of trust were also defined to better define trust in an AI-AMD system. An important concept of trustworthiness was also highlighted in this section to clarify the differences between trust and trustworthiness.

1. Default Trust, Impersonal Trust and Human-AI Trust

Often, humans do not pause to deliberately ponder about whether or how they trust the machine or technology they use. This is called default trust. This happens when humans accept the decisions and actions of machines without any concerns. On the other hand, the notion of trust is very deliberative. In such cases, people exercise extreme caution when deciding whether to trust a machine or technology, as well as the conditions connected with trusting machines (Hoffman et al. 2013).

Trust is the foundation of most of the fundamental relationships in human life; and without it, many valuable social interactions and bonds would be jeopardized. Without trust, we are going to be isolationist and suspicious because of fear of harm and dishonesty (O’neill 2002). To trust someone often requires a belief about their trustworthiness. Trusting someone can be about having confidence in them to carry out a particular task. To be trustworthy, one must have the ability to satisfy the trust that is placed on them. One must be capable of being trusted (Ryan 2020).

There are several definitions of trust in the literature. According to a review of the literature, trust can be described as a belief, attitude, intention, or behavior. Fishbein and Ajzen (1977) proposed a model to help to reconcile the conflicting nature of the definitions of trust as depicted in Figure 3. Based on the authors' model, behavior is a result of intention; and intention is derived from attitudes. Beliefs and perceptions serve as the foundation for trust as well as knowledge base that influences attitudes. The amount and accuracy of information, as well as human encounters, impact people's beliefs.

Attitude refers to an emotional assessment of beliefs that motivates humans to act in a certain way (Lee and See 2004). Intention then leads to behavior, which is influenced by the cognitive, social, and external environment restrictions. Within the context of trust, Lee and See (2004) argued that trust is an attitude. A person's beliefs create trust, and different levels of trust will lead to different intentions and behaviors. As such, Lee and See (2004) defined trust as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Thus, trust aims to eliminate ambiguity in a person's life by anticipating the outcome of specific conduct, and to make it easier for individuals to work together (Jacovi et al. 2021).

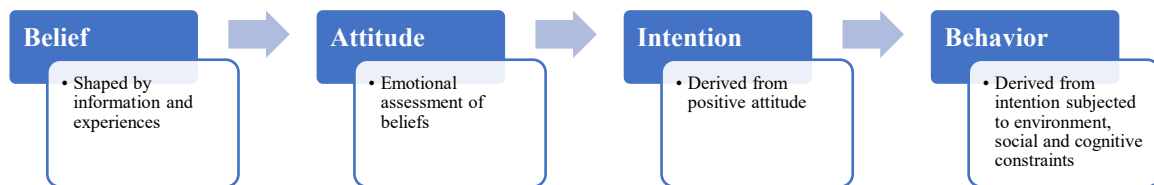


Figure 3. Framework of Trust. Adapted from Ajzen and Fishbein (1977).

Like interpersonal trust, human-AI trust has been a critical element of the HMI, as incorrect levels of trust will result in undesirable consequences. According to Jacovi et al. (2021) and Hoffman (2017), trust between humans and AI systems includes the concepts of anticipation, uncertainty, and vulnerability. Humans want to develop interpersonal trust between each other to reduce uncertainty and risk of working together. This is supported

by the expectation or conviction that one side will act in the best interests of the other. Trust occurs between humans and AI for the same reason.

2. Contractual Trust

Human-AI trust is premised on what the human anticipates in the AI system's behavior. Trust is not built based on the general AI system performance ability, but the specific tasks with explanation that a human expects from the AI system (Hawley 2014; Tallant 2017). This is contractual trust. Human-AI contractual trust happens when a person believes that the AI system will follow a specific contract established between them. Contract can refer to any features that the AI system performs even if it is not the original performance function that the AI system was trained for. As contracts specify the behavior anticipated from an AI system, they become important in understanding human-AI trust as it enables us to make clear the objective of anticipation. As such, to trust an AI system means that human believes that contracts will be maintained and accepts vulnerability to the AI's actions (Jacovi et al. 2021). Then, by trusting the AI system, humans hope to predict that the AI system will uphold the contract in the face of uncertainty.

3. Trust Repair

“Once lost, trust in automation, like interpersonal trust, can be hard to re-establish.”

— Hoffman et al. (2013,85)

Another important aspect of trust - trust repair - is defined an action to improve trust after a trust violation has taken place. Researchers de-Visser, Pak, and Shaw (2018) demonstrated that being able to repair trust contributes to a healthy relationship. For example, the best functioning human teams are those who participate in team building, trust rebuilding, and exposing one's own emotional vulnerabilities (Duhigg 2016). High performing human team members will engage in activities to help each other perform better for a better team performance. This is the type of behavior we need to implant in AI systems in order to create a high-performing human-AI collaboration.

However, numerous researchers have demonstrated that there are significant differences between humans and AI systems. This is because AI systems are perceived to

be more reliable due to their invariance in performance. Any trust repairs actions such as apology notifications or explanations might be less effective if humans perceive that the AI systems behaviors are fixed and not likely to change (de Visser et al. 2016). Additionally, AI systems performing trust repair actions such as apology prompts, or explanations may be seen by humans as inadequate, as autonomous systems are expected to change their behaviors and learn from their errors. Hence, this aspect of trust repair on autonomous systems will be considered in this thesis to ensure that trust can continue to improve for sustained performance (de Visser et al. 2016).

4. Trustworthy AI

In human-AI trust, trust and being trustworthy (trustworthiness) are distinctly different. Trust as defined earlier is an attitude of the human operator while trustworthy is a feature of the AI system. An AI system is considered trustworthy to some contracts if it can uphold the contracts that the operator desires. Trust and trustworthiness are independent because the pursuit of one may not necessarily lead to the other. Trustworthiness may not always lead to trust. On the contrary, trust in an AI system can exist even with an untrustworthy system (Jacovi et al. 2021; Wright 2010). Lee and See (2004) and Jacovi et al. (2021) asserted that warranted trust exists because of trustworthiness. Else, unwarranted trust is being developed. In other words, warranted trust happens when the anticipation of the desired AI system behaviors becomes a reality. To develop warranted trust, there is a need to influence the trustworthiness of the AI model in the system.

5. Trust Calibration, Resolution and Specificity

Trust is a crucial element of human-AI interactions, since improper levels of trust may result in the AI system being misused or disused (Wicks, Berman, and Jones 1999). Misuse of AI system occurs when there is over-reliance on the system which may cause human monitoring failures or decision biases. Disuse of AI system occurs when humans neglect or underutilize the system (Parasuraman and Riley 1997).

Establishing the “right” amount of trust is critical in preventing AI system misuse and abuse since it promotes productive human-AI relationships (Wicks, Berman, and Jones

1999). The ideas of trust calibration, trust resolution, and trust specificity explain the gaps between trust and AI capabilities (Lee and See 2004). Trust calibration is the link between human-AI trust and AI capabilities (Lee and Moray 1992). When human places more trust than the what the AI system is capable of, over-trust (a poor calibration) happens. This will lead to misuse, in which a person is overly dependent on AI, failing to overrule the AI system when needed. On the other hand, distrust happens when the human-AI is falls below of that AI system capabilities. This leads to disuse in which a person fails to use AI when doing so would enhance the overall performance. As a result, a person's trust in the AI system must be calibrated so that acceptable levels of human-AI trust grow to prevent misuse or disuse (Merritt et al. 2013; Parasuraman and Miller 2004). The 45-degree line in Figure 4 depicts good calibration of trust (calibrated trust), where the amount of trust matches with AI capabilities. Over-trust is represented in the region above the diagonal line and below it is distrust.

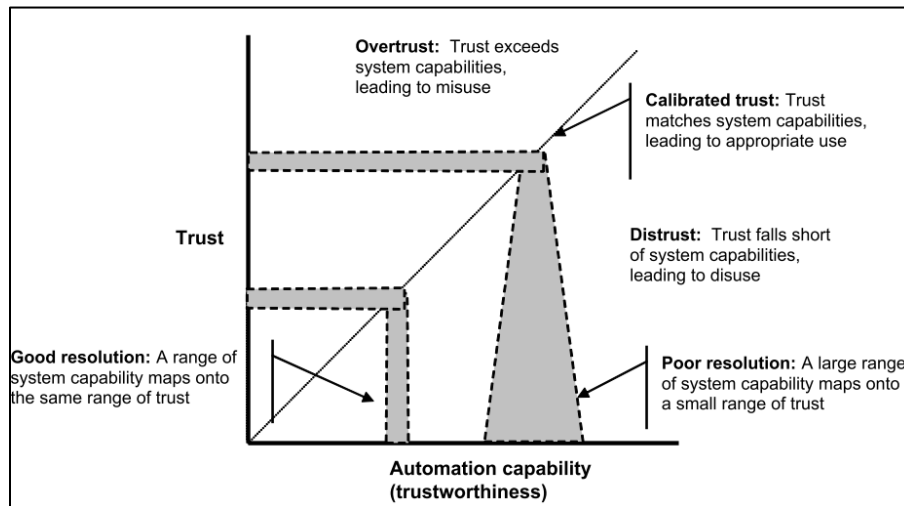


Figure 4. Relationship between Trust Calibration and Automation Capability. Source: Lee and See (2004).

Trust resolution is defined as the precision of a trust judgement that distinguishes the levels of automation (LOA) capability (Cohen, Parasuraman, and Freeman 1998). Figure 4 shows that poor resolution is a result of having a wide range of AI capability being

mapped onto a small range of trust. When trust resolution is poor, the human-AI trust is less reflective of the AI capabilities related because huge changes in automation capability do not bring about the same level of changes in trust.

Functional specificity refers to the extent to which trust is reflective of the functions, sub-functions, or modes of the AI. A high level of functional specificity in an AI system indicates that an individual's confidence reflects a specific sub-function capacity or mode of AI. Low functional specificity, on the other hand, suggests that an individual's trust reflects the AI system's overall capabilities (Lee and See 2004).

Specificity can also describe variations in trust caused by varying situation or over time. Researchers of trust in AI generally acknowledge that trust is dynamic rather than being defined by a state. Trust increases or declines over time (Hoffman 2017). The sensitivity of trust to changes in AI capability caused by environmental changes is described by temporal specificity. A person's trust in a situation with great temporal specificity reflects the ongoing changes in AI capability throughout time. On the other hand, because trust has a low temporal specificity, it only reflects long-term differences in AI capabilities. Temporal specificity provides insights on the improvements to implement when the environment changes, affecting the AI system's capabilities.

In conclusion, calibrated trust, good trust resolution, and high specificity can decrease the possibilities of AI misuse and disuse. Trust with a high functional and temporal specificity suggests that the changes in trust will likely be influenced by the changes in the AI capabilities over time (Lee and See 2004). These concepts can guide the engineering design, assessment, and training to improve human-AI partnerships.

C. HUMAN-MACHINE INTERACTIONS IN ARTIFICIAL INTELLIGENCE SYSTEMS

This section contains information gathered about existing work on human-machine interactions (HMI). The literature review included a review of HMI concepts and characteristics, as these HMI will influence human trust in a system (Abbass 2019).

The review of HMI included HMI models, levels of automation corresponding to human information processing, mappings of system functions that can be automated, and

HMI design considerations and constructs. A case study was reviewed that proposes an automation taxonomy for air traffic management. The case study was reviewed to draw additional insights for effective HMI.

1. Characterization of Human-Machine Interactions

In this thesis, “machine” is regarded as an AI system. HMI is defined as the interactions and communications between people and autonomous systems via a user interface (Toreini et al.). According to Norman (1984), HMI can be characterized into a series of four stages of activities: (1) intention, (2) selection, (3) execution, and (4) evaluation as shown in Figure 5.

- Intention stage is the human’s mental interpretation of a goal.
- Selection stage translates the intention into actions. Humans must review all possible options and determine the best option that satisfies the intention. Then, the actual sequence of actions must be specified to the machine.
- Execution is the stage where humans physically enter the information into the computer to carry out the tasks.
- Evaluation is the stage where the results of the actions are reviewed for further activity. Feedback is sought to improve the system via an iterative loop.

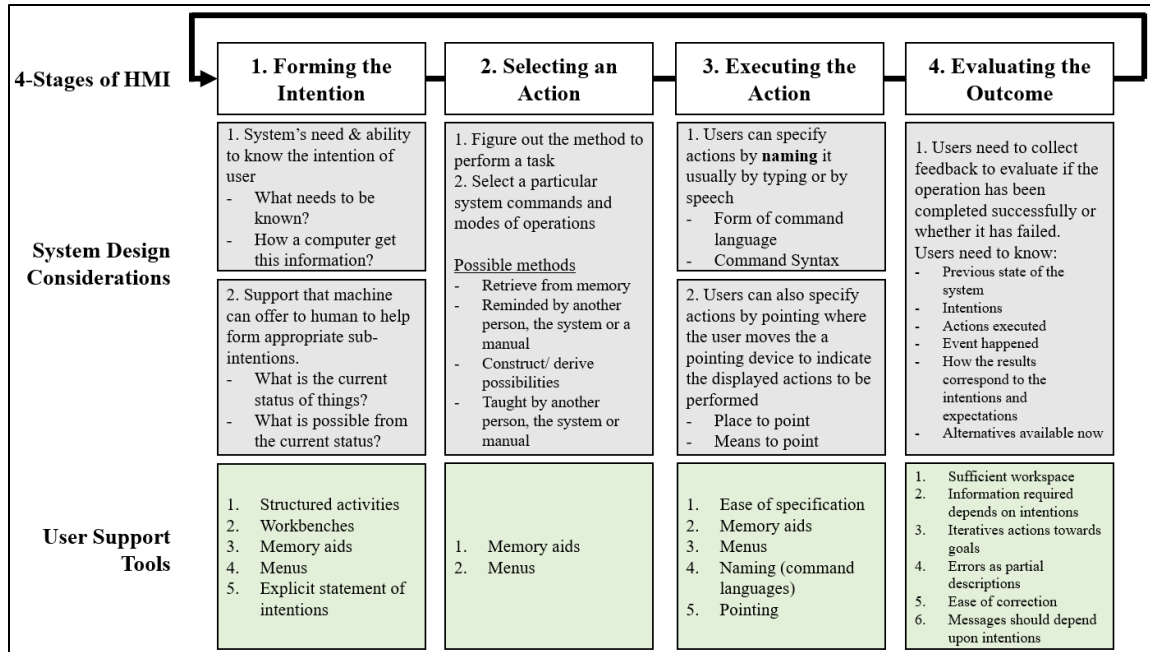


Figure 5. Stages, Design Considerations and Support Tools of HMI Activities. Adapted from Norman (1984).

This model of HMI characterization is just an approximation of the actual interactions between humans and machines. Humans may not possess or execute well-defined stages of decision and action. Humans are often not fully aware of the reasons for their own actions; humans sometimes perform actions that do not fulfil their intentions. Humans' decision-making processes are comprised of complicated and highly parallel processing that includes both conscious and sub-conscious processing of many factors. Nevertheless, Norman's (1984) four-stage series model approximations provide relevant insights in this thesis.

The system design considerations relevant to each of the stages of HMI are also summarized in Figure 5. These considerations serve as guides to assist designers in building their systems according to the four stages of HMI to enhance the interactions between operators and the automated system. Additionally, the user support tools relevant to each stage that can help to improve the HMI are summarized at the bottom of Figure 5. These tools help to address the system design considerations to improve HMI.

2. Levels of Machine Automation and Human Information Processing – Deep Dive into “Selecting an Action” Stage of Human-Machine Interactions

Automation is the partial or complete replacement of a task or function formerly carried out by a human (Parasuraman, Sheridan, and Wickens 2000). This concept implies that a system’s level of automation (LOA) does not have to be “all or nothing.” The LOA of a system can range from entirely manual to completely automatic. Table 1 shows a 10-point scale for automation, with higher values indicating a higher autonomy of the system over human action. For instance, at lower levels of automation, the machine recommends courses of actions, but the human has the ultimate authority for executing the courses of action (COA). On the contrary, at higher levels of automation, the machine has more control over decisions and actions, leaving humans with simpler roles.

Table 1. 10-point Scale for Automation of Decision and Action Selection.
Source: Parasuraman, Sheridan, and Wickens (2000).

HIGH	10. The computer decides everything, acts autonomously, ignoring the human.
	9. informs the human only if it, the computer, decides to
	8. informs the human only if asked, or
	7. executes automatically, then necessarily informs the human, and
	6. allows the human a restricted time to veto before automatic execution, or
	5. executes that suggestion if the human approves, or
	4. suggests one alternative
	3. narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
LOW	1. The computer offers no assistance: human must take all decisions and actions.

Sheridan (2002) also proposed the notions of shared control and control trading. When a human and a machine control various components of the system simultaneously, this is referred to as shared control. The machine is employed in a shared control system to

augment humans' abilities beyond what humans can do or to assist humans in making their work easier (Sheridan 2002). On the other hand, trading of control happens when the machine is employed to take over the role of the human or backs up the human when the human falters. In addition, Sheridan (2002) also commented that there are forms of cooperative control which happens when the control is triggered by either the human or machine and the other processes or improves it.

The key challenge in a shared control system is proper distribution of tasks between humans and machines to ensure coherence actions for both the human operators and machines. The key challenge in a trading control system is determining when control should be handed over to the machine and when it should be returned to the human.

Parasuraman, Sheridan, and Wickens (2000) proposed an extended form of the model to include the automation of various types of functions such as input functions present in HMI. Table 1 shows the levels of automation, which largely apply to the automation of output functions such as decisions and actions. As such, a four-stage model representing human information processing was developed as shown in Figure 6.

- Stage 1 – Sensory processing is the receiving and registering of information from different sources. This stage involves the placement of sensors, sensory processing, handling of data, understanding of data information, and selective attention.
- Stage 2 – Perception/working memory involves mindful view and understanding of the retrieved and processed information that is available in the human mind. This stage consists of cognitive processes like inference, assimilation and rehearsal that are carried out before the decision point.
- Stage 3 – Decision making is the cognitive processing of making decisions
- Stage 4 – Response selection is to act upon the decision made.

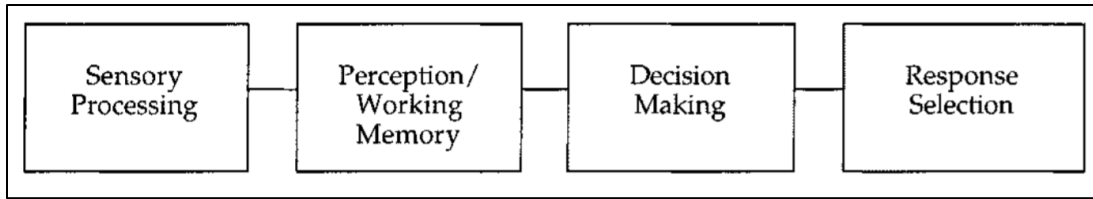


Figure 6. Four-Stage Model of Human Information Processing. Adapted from Parasuraman, Sheridan, and Wickens (2000).

Parasuraman, Sheridan, and Wickens (2000) has mapped the above four-stage model to equivalent system functions that may be automated as shown in Figure 7. The different LOA as depicted in Table 1 can be applied to these system functions.

- Function 1 – The gathering, identification, and recording of input data is information acquisition. These tasks are analogous to the stage of human information processing.
- Function 2 – Information analysis involves cognitive functions like discovering and interpreting information. At high levels of automation, it involves fusing of multiple input data into a single data.
- Function 3 – Decision and action selection involve making a choice among various options. Different degree of substitution or augmentation of human selection process can be automated with machine processes.
- Function 4 – Action implementation translates the COA into executable actions. At this stage, automation entails varying degrees of machine execution that substitute human hands or voices.

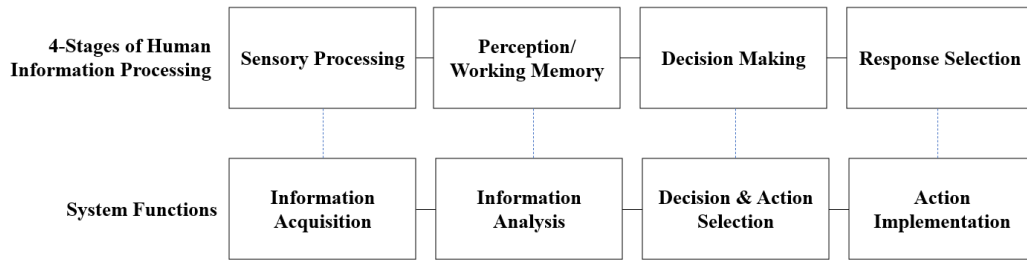


Figure 7. Mapping of System Functions to Four-Stage Model of Human Information Processing. Source: Parasuraman, Sheridan, and Wickens (2000).

It is also worth noting that the LOA for these functions can change based on the system context (situation and environment) during actual operational use. Figure 8 depicts a framework developed by Parasuraman, Sheridan, and Wickens (2000) for automation design. These authors' framework proposes a stepwise and iterative process to guide automation design; in determining whether automation should be utilized to a particular function. Next, the level of automation of each function should be determined. In this model, the authors proposed that the level of automation should be assessed by first considering human limitations such as cognitive workload, situational awareness (SA), and skills decay. Then, secondary factors such as automation capability, and associated impacts of automation decision/action consequences should be considered. The application and evaluation of these factors should be an iterative process too. However, it is noted that this framework is not a "be-all and end-all," but it serves as a useful guiding principle for subsequent detailed consideration.

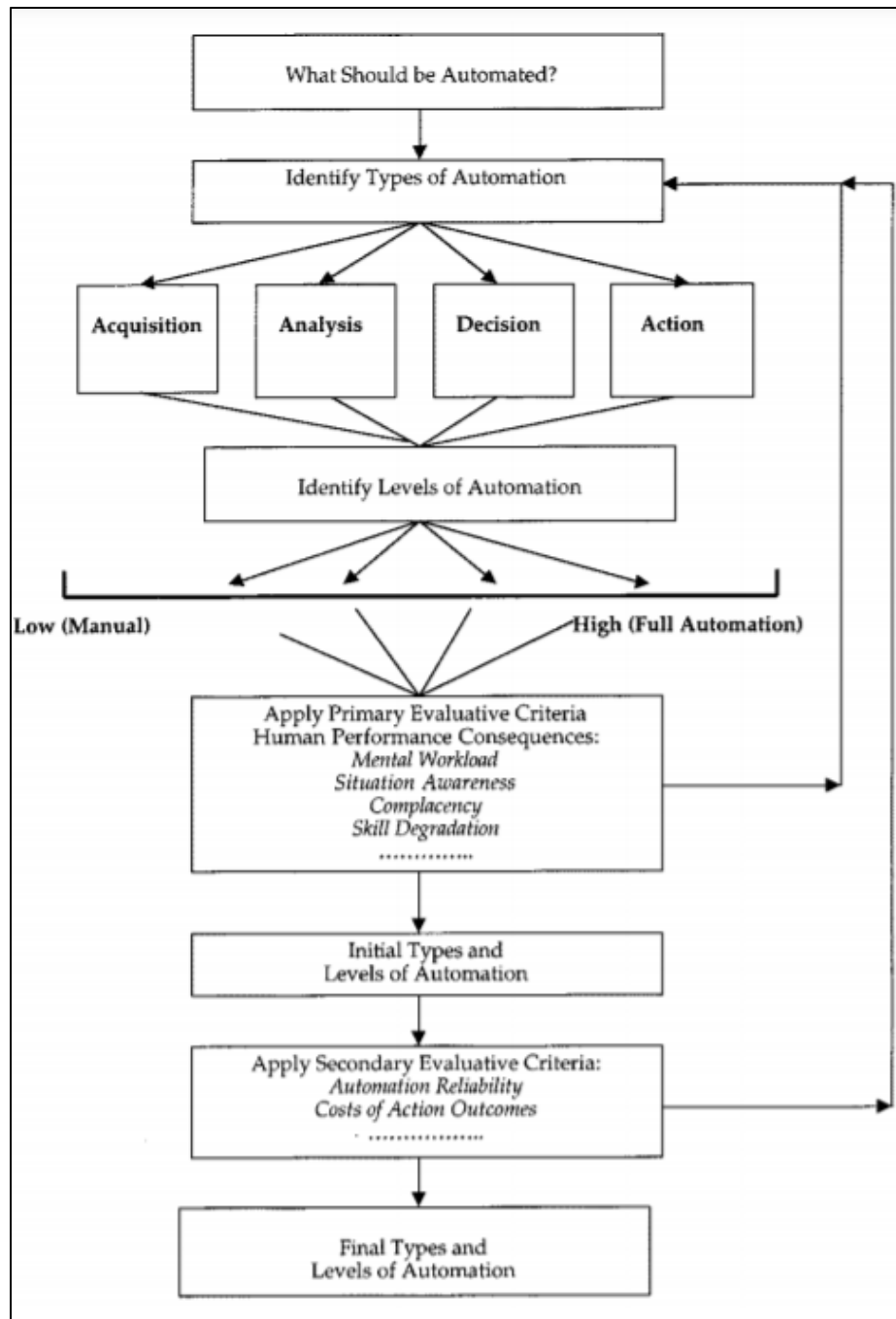


Figure 8. Flow Chart to Guide Levels of Automation Design. Source: Parasuraman, Sheridan, and Wickens (2000).

3. Human-Machine Interactions Design Constructs

Pritchett and Feary (2011) proposed that to describe HMI sufficiently, five constructs of authority, teams, work, representations, and interface mechanisms must be established in a coordinated manner at varying levels of abstraction as shown in Figure 9.

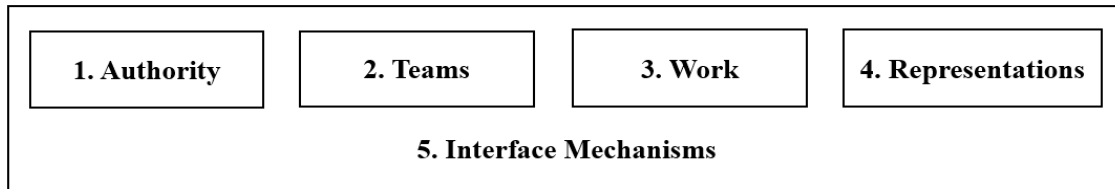


Figure 9. Five Constructs of HMI. Adapted from Pritchett and Feary (2011).

a. Construct 1 – Authority

Authority is defined as the accountability of outcomes that is demanded of an agent who is performing it. At the highest level of abstraction, an authority on the automation agent must be defined for accountability for outcomes. When accountability for an outcome is demanded of an automation agent with the power to execute it, the true capability exists. The difficulty for true authority is to ensure that accountability for all outcomes to agents is clearly defined throughout the lifetime of operations. Hence, establishing effective HMI requires coherent technology design (machine capabilities and methods of execution) and policy (accountability).

b. Construct 2 – Team

A team is defined as individuals who work together inter-dependently to accomplish a shared goal. A construct about *teams* is to examine how the humans and automation systems coordinate their specific functions like shared situation awareness (SA) and team maintenance. Autonomous systems are regarded as teammates. A shared mental model, often known as “team knowledge” or “shared situation awareness,” underpins teamwork. Conditions for a shared mental model include: (1) each possesses the requisite knowledge, (2) each knows what information the rest possess should he seeks it,

and (3) each knows which information and when the other teammate needs it. The mental model supports team interactions and assists each team member create a clear expectation of each other and task. As such, achieving a shared mental model for teams is a common goal of many training programmes.

Shared mental models amongst teammates are also important for coordination and communication for high performing teams. However, communications amongst teammates can be disrupted. As a result, it is critical to “build in” or engineer team structures and shared mental models that enable teammates to anticipate one another’s information needs and supply information at non-interruptive junctures. However, the inclusion of automation demands additional requirements for communications and synchronization amongst teammates. So, the challenge for effective HMI is to integrate the essence of human-machine teams interactions into action-oriented guides for automation designers to better assist and fulfil the needs of humans.

c. Construct 3 – Work

Work engages in purposeful activity. A construct about work is to examine the activities and interactions required across all automation agents to perform the tasks and meet their goals. It is important to address the segregation of work in terms of “what needs to be done” and “who does what” to meet the system goals. Beyer (1998) developed a contextual design model of work to design information technology for somewhat static domains. This model has five perspectives as follows:

- Data and items exchanged amongst teammates
- Organizational culture influencing work
- Physical environment supporting work
- Support tools for work
- Sequence of activities to enable work

This model can be extended to work in a dynamic environment by including analysis on mission task, purpose, work objectives, success metric, utility function, event

triggers, task cruciality, timing, and limiting conditions for the work processes (Billman et al. 2010).

d. Construct 4 – Representation

Representation is defined as the presentation within the mind of an image, idea, or concept. A construct about representation is to examine how humans perceive and conceptualize the environment to design in automation so that its actions are understandable to, and interactable with, humans. External representations, such as interfaces and displays, according to Zhang and Norman (1994), can influence cognitive behavior, change the character of human tasks, and give easily understood information that can be used easily. Hence, these representations seek to improve or leverage human perceptual and cognitive processes to improve performance. Pritchett and Feary (2011) argued that effective HMI is dependent on shared representations except for instances where the automated system can execute its tasks without any inputs and supervision from humans. Specifically, HMI involves these four representations:

- The automated system must have an internal representation of the functions such as controls loops, event triggers or established rules to carry out the work.
- Humans must also possess an internal representation of cooperative work. The representation will be the most elaborated for the exact work an operator is assigned to. However, it is important for the operator to have a representation of their teammates' work to understand their requirements or activities for effective teaming.
- An explicit representation of the automation's output and processes must be displayed to the human. Also, some automated systems may display humans' actions and processes.
- A representation of the environment may also be made available to the human.

These four representations suggest that there is a need to make sure that there is observability of the automated system's functioning, as well as coherent internal and external representations that drive the system's behavior. Additionally, the automation's functioning should be understandable to the human, and consistent with the work

environment. As such, a common representation construct between humans and machines is needed.

e. Construct 5 – Interface Mechanisms

The interface mechanisms construct investigates the exchange of information between humans and automated systems, and the actions required by each entity to carry out planned behavior. Interfaces are generally where the problems with HMI are revealed. Current interface technology such as keyboards, computer displays, verbal communication, gesture communication, buttons, and other input tools to input instructions or commands all have their own limitations and present HMI issues. Nonetheless, interface difficulties can be addressed with integrated tools by considering the four constructs outlined above with the characteristics of the interfaces and the way human processes information. This HMI design should allow designers to (Pritchett and Feary 2011):

- Improve understanding of both targeted performance measurements and human preferences indicating competing performance trade-offs in the work domain.
- Transform what has been learnt about the job into a suitable representation.
- Design interfaces by matching activity representation to suitable interface devices.
- Employ human performance analysis and computational automation tools to assess prototypes during system development and optimize the interface performance against the performance metrics and human preferences.

4. Joint Activity Theory of Human-Machine Interactions

Bradshaw, Feltovich, and Johnson (2012) argued that the HMI is not so much about the allocation of tasks to humans or machines, but rather should focus on how tasks are allocated and shared between humans and machines who work together. For instance, Hoffman et al. (2002) developed a list to demonstrate how humans and machines can complement each other through mutual interactions to improve the overall performance as shown in Table 2. Understanding both human and machine strengths and limitations is

crucial for effective HMI to bring them together to enhance the performance of human-machine teams.

Table 2. HMI to Enhance Performance. Adapted from Bradshaw, Feltovich, and Johnson (2012).

Machines	
Are constrained in that:	Need people to:
Sensitivity to context is low and is ontology-limited	Keep them aligned to context
Sensitivity to change is low and recognition of anomaly is ontology-limited	Keep them stable given the variability and change inherent in the world
Adaptability to change is low and is ontology-limited	Repair their ontologies
They are not “aware” of the fact that the model of the world is itself in the world	Keep the model aligned with the world
People	
Are not limited in that:	Yet they create machines to:
Sensitivity to context is high and is knowledge- and attention-driven	Help them stay informed of ongoing events
Sensitivity to change is high and is driven by the recognition of anomaly	Help them align and repair their perceptions because they rely on mediated stimuli
Adaptability to change is high and is goal-driven	Effect positive change following situation change
They are aware of the fact that the model of the world is itself in the world	Computationally instantiate their models of the world

Bradshaw, Feltovich, and Johnson (2012) went on to examine successful HMI using joint activity theory. Joint activity theory focuses on interdependence where parties come together to deliver a genuine joint product. The authors summarized the following maxims for an effective HMI as shown in Table 3.

Table 3. Eight Maxims for Good Human or Machine Agent. Source: Bradshaw, Feltovich, and Johnson (2012).

1.	A good agent is <i>observable</i> . It makes its pertinent state and intentions obvious.
2.	A good agent is attuned to the requirement of <i>progress appraisal</i> . It enables others to stay informed about the status of its tasks and identifies any potential trouble spots ahead.
3.	A good agent is <i>informative</i> and <i>polite</i> . It knows enough about others and their situations so that it can tailor its messages to be helpful, opportune, and appropriately presented.
4.	A good agent <i>knows its limits</i> . It knows when to take the initiative on its own, and when it needs to wait for outside direction. It respects policy-based constraints on its behavior, but will consider exceptions and workarounds when appropriate.
5.	A good agent is <i>predictable</i> and <i>dependable</i> . It can be counted on to do its part.
6.	A good agent is <i>directable</i> at all levels of the sense-plan-act cycle. It can be retasked in a timely way by a recognized authority whenever circumstances require.
7.	A good agent is <i>selective</i> . It helps others focus attention on what is most important in the current context.
8.	A good agent is <i>coordinated</i> . It helps communicate, manage, and deconflict dependencies among activities, knowledge, and resources that are prerequisites to effective task performance and the maintenance of “common ground.”

5. Automation Taxonomy for HMI in Air Traffic Management

Save, Feuerberg, and Avia (2012) presented a new taxonomy of autonomy to identify and compare various types of automation that help in air traffic management. However, Save, Feuerberg, and Avia (2012) argued that the principles and the description of this taxonomy are not only limited to this field of air traffic management but transferable to other domains. This taxonomy is grounded on Sheridan and Verplank (1978) idea to decide on the extent of automation rather than whether to automate a task entirely or not and Parasuraman, Sheridan, and Wickens (2000) framework of the four generic HMI functions. While the earlier models succeeded in acknowledging the different nature of automation support, specific levels for each of the functions have not been defined. This reduced the model’s ability to provide human aspects recommendations in automation design. As such, this taxonomy attempted to address the issue of identifying different LOA by considering the various ways in which automation could increase human performance.

The taxonomy was developed based on the following principles and are given in Tables 4–6:

- An automated system cannot have one “overall” level of automation applied to all functions. Different levels of automation must be applied to specific functions within the system.
- One automated system can support more than one function. Each function is expected to have different LOA.
- The description of each automation level takes human performance factors into consideration.

Table 4. Proposed New Taxonomy for Air Traffic Management (Part I).
Source: Save, Feuerberg, and Avia (2012).

A INFORMATION ACQUISITION	B INFORMATION ANALYSIS	C DECISION AND ACTION SELECTION	D ACTION IMPLEMENTATION
A0 Manual Info Acquisition	B0 Working Memory Based Info Analysis	C0 Human Decision Making	D0 Manual Action and Control
The human acquires relevant information on the process s/he is following without using any tool.	The human compares, combines and analyses different information items regarding the status of the process s/he is following by way of mental elaborations. S/he does not use any tool or support external to her/his working memory.	The human generates decision options, selects the appropriate ones and decides all actions to be performed.	The human executes and controls all actions manually.
A1 Artefact-Supported Info Acquisition	B1 Artefact-Supported Info Analysis	C1 Artefact-Supported Decision Making	D1 Artefact-Supported Action Implementation
The human acquires relevant information on the process s/he is following with the support of low-tech non-digital artefacts. <i>Ex. 1) Identification of aircraft positions on an aerodrome/airport according to Procedural Air Traffic Control rules and without use of radar support.</i>	The human compares, combines, and analyses different information items regarding the status of the process s/he is following utilising paper or other non-digital artefacts. <i>Ex. 1) Use of flight strips to compare altitudes/levels/pl. times of different aircraft and to pre-plan future traffic.</i>	The human generates decision options, selects the appropriate ones and decides all actions to be performed utilising paper or other non-digital artefacts.	The human executes and controls actions with the help of mechanical non-software based tools. <i>Ex. 1) Use of a hammer or leverage to increase the kinetic energy of human gesture. Ex. 2) Use of a mechanical or hydraulic rudder to achieve a change in direction.</i>
A2 Low-Level Automation Support of Info Acquisition	B2 Low-Level Automation Support of Info Analysis	C2 Automated <u>Decision Support</u>	D2 Step-by-step Action Support:
The system supports the human in acquiring information on the process s/he is following. Filtering and/or highlighting of the most relevant information are up to the human. <i>Ex. 1) Identification of aircraft positions in the airspace by way of Primary Radar working positions.</i>	<u>Based on user's request</u> , the system helps the human in comparing, combining and analysing different information items regarding the status of the process being followed. <i>Ex. 1) Activation by ATCOs of Speed Vectors for specific tracks on the CWP, in order to anticipate potential conflicts in a defined time frame.</i>	The system proposes one or more decision alternatives to the human, leaving freedom to the human to generate alternative options. The human can select one of the alternatives proposed by the system or her/his own one. <i>Ex. 1) AMAN visualization of the proposed sequence of aircraft.</i>	The system <u>assists</u> the operator in performing actions by executing part of the action and/or by providing guidance for its execution. However, each action is executed based on <u>human initiative</u> and the human keeps full control of its execution. <i>Ex. 1) The aural and visual component of TCAS RA in current TCAS II version 7.0 (also LOA C5)</i>

Table 5. Proposed New Taxonomy for Air Traffic Management (Part II).
Source: Save, Feuerberg, and Avia (2012).

A3 Medium-Level Automation Support of Info Acquisition	B3 Medium-Level Automation Support of Info Analysis	C3 Rigid Automated Decision Support	D3 Low-Level Support of Action Sequence Execution
The system supports the human in acquiring information on the process s/he is following. It helps the human in <u>integrating</u> data coming from different sources and in <u>filtering</u> and/or <u>highlighting</u> the most relevant information items, <u>based on user's settings</u>. <i>Ex. 1) CWP allowing ATCOs to set flight level filters to display only certain traffic on the screen.</i>	<u>Based on user's request</u>, the system helps the human in comparing, combining and analysing different information items regarding the status of the process being followed. The system <u>triggers visual and/or aural alerts</u> if the analysis produces results requiring attention by the user. <i>Ex. 1) ERATO Filtering and What-if function.</i> <i>Ex 2). VERA Tool to display the closest point of approach between two aircrafts.</i>	The system proposes one or more decision alternatives to the human. The human can only select one of the alternatives or ask the system to generate new options.	The system performs automatically a sequence of actions <u>after activation by the human</u>. The human maintains full control of the sequence and can modify or interrupt the sequence during its execution. <i>Ex. 1) Explicit initiation of an electronic coordination with adjacent sector via digital input (replacing use of telephone).</i>
A4 High-Level Automation Support of Info Acquisition	B4 High-Level Automation Support of Info Analysis	C4 Low-Level Automatic Decision Making	D4 High-Level Support of Action Sequence Execution
The system supports the human in acquiring information on the process s/he is following. The system <u>integrates</u> data coming from different sources and <u>filters</u> and/or <u>highlights</u> the information items which are considered relevant for the user. The <u>criteria</u> for integrating, filtering and highlighting the relevant information are <u>predefined at design level</u> but <u>visible to the user</u>. <i>Ex. 1) D-TAXI tool (including graphical route information)</i>	The system helps the human in comparing, combining and analysing different information items regarding the status of the process being followed, based on parameters pre-defined by the user. The system <u>triggers visual and/or aural alerts</u> if the analysis produces results requiring attention by the user. <i>Ex. 1) MTCD visual alerts (allowing some tuning of parameters by the user)</i>	The system generates options and decides autonomously on the actions to be performed. The human is informed of its decision. <i>Ex. 1) Aural and visual component of TCAS RA in current TCAS II version 7.0 (also LOA D2)</i>	The system performs automatically a sequence of actions <u>after activation by the human</u>. The human can <u>monitor</u> all the sequence and can <u>interrupt</u> it during its execution. <i>Ex. 1) Acknowledgment by pilot of a clearance received through CPDLC (data-link) and automatically sent to FMS and autopilot.</i> <i>Ex. 2) Autopilot following the FMS trajectory.</i>

Table 6. Proposed New Taxonomy for Air Traffic Management (Part III).
Source: Save, Feuerberg, and Avia (2012).

A5 Full Automation Support of Info Acquisition	B5 Full Automation Support of Info Analysis	C5 High-Level Automatic <u>Decision Making</u>	D5 Low-Level <u>Automation</u> of Action Sequence Execution
The system supports the human in acquiring info on the process s/he is following. The system <u>integrates</u> data coming from different sources and <u>filters</u> and/or <u>highlights</u> the information items considered relevant for the user. The <u>criteria</u> for integrating, filtering and highlighting are <u>predefined at design level</u> and <u>not visible to the user</u>	The system performs comparisons and analyses of data available on the status of the process being followed <u>based on parameters defined at design level</u> . The system <u>triggers visual and/or aural alerts</u> if the analysis produces results requiring attention by the user. <i>Ex. 1) STCA visual and aural alerts.</i>	The system generates options and decides autonomously on the action to be performed. The human is informed of its decision only on request. (Always connected to to an <i>Action Implementation</i> level not lower than D5.)	The system <u>initiates and executes</u> automatically a sequence of actions. The human can <u>monitor</u> all the sequence and can <u>modify or interrupt</u> it during its execution. <i>Ex. 1) Implicit initiation of an electronic co-ordination with adjacent sector as agreed exit conditions (according to Letter of Agreement) cannot be met anymore after changes to the a/c trajectory has been made.</i>
		C6 Full Automatic <u>Decision Making</u>	D6 Medium-Level <u>Automation of Action</u> Sequence Execution
		The system generates options and decides autonomously on the action to be performed without informing the human. (Always connected to an <i>Action Implementation</i> level not lower than D5.)	The system <u>initiates and executes</u> automatically a sequence of actions. The human can <u>monitor</u> all the sequence and can <u>interrupt</u> it during its execution. <i>Ex. 1) TCAS AP/FD during a corrective TCAS RA.</i>
			D7 High-Level <u>Automation of Action</u> Sequence Execution
			The system <u>initiates and executes</u> a sequence of actions. The human can only <u>monitor part of it</u> and has <u>limited opportunities to interrupt it</u> .
			D8 Full <u>Automation of</u> Action Sequence Execution
			The system <u>initiates and executes</u> a sequence of actions. The human cannot monitor nor interrupt it until the sequence is not terminated.

D. TRUST FACTORS IN AI SYSTEMS

This section offers an overview of the elements that influence the establishment of human trust in automated/autonomous systems. The review of the trust factors is critical to deriving strategy and engineered solutions to improve the trust between humans and AI-AMD systems. Broadly, three schools of thought on trust developments were reviewed:

First, Schaefer et al. (2016) proposed a three-factor trust model that is based upon the human, the partner (system), and the environment of trust. It is a meta-analysis of other researchers' work on the trust development of automated systems. Other researchers' work was also reviewed to complement the analysis done by Schaefer et al. (2016).

Second, Sheridan (2019) categorized trust factors into objective and subjective attributes. His work emphasized the importance of subjective factors in influencing trust development in current and future automated and autonomous systems. These subjective factors emerged against the backdrop of more advanced and intelligent automation and autonomous systems.

Third, Denning (2016) developed a six-user level software quality assessment. Denning's work was based on the observation that trust is made by human beings. While there could be objective measures to support trust development, there is no algorithm that can make the final determination of trust. It is ultimately a human's decision. As such, he proposed an assessment metric that addresses the fundamental questions of whether the software helps the operators in their work and whether the operator can depend on it. These insights further support Sheridan's (2019) assertion on the emergence of subjective factors.

1. Schaefer et al.'s (2016) Three-Factor Trust Model

Following an exhaustive analysis of trust in automation systems, Schaefer et al. (2016) suggested a three-factor trust model. In this section of the thesis, the Schaefer et al. (2016) model will be adopted as the main framework for the analysis of trust factors and supplemented with other researchers' work. Figure 10 depicts the researchers' three-factor trust model. The trust factors are broadly classified as (1) human-related, (2) partner-related, and (3) environment-related.

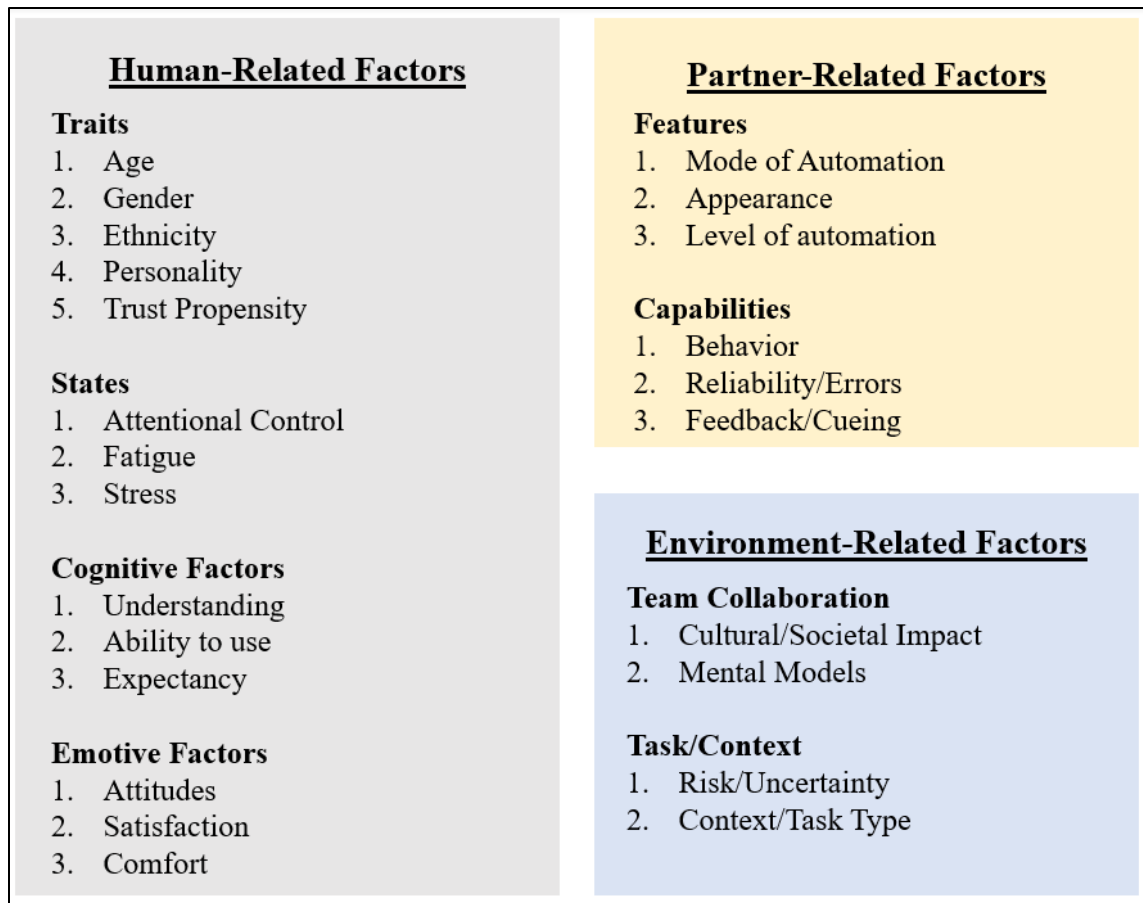


Figure 10. Three-factor Trust Model. Adapted from Schaefer et al. (2016).

a. Humans-Related Factors of Trust

The study of trust between humans and machines should begin with human-related factors. Arguably, the character and the nature of the human operator could potentially exert the most significant influence on the overall performance of the system (Schaefer et al. 2016). The factors associated with humans are (a) human traits, (b) states, (c) cognitive factors, and (d) emotive factors.

(1) Traits

Humans' traits and the demographic characteristics they are exposed to can affect how they interact with automated systems. These traits consist of age, gender, ethnicity, and personality. However, it was found out that there was no consistent relationship across

different studies. The link was discovered to be context-specific and to change depending on the parameters of the automated task (Szalma and Taylor 2011).

- Age – According to studies, elderly persons appear to rely more on automation than younger adults. As a result, they are more relaxed with AI systems (Ho, Wheatley, and Scialfa 2005). It was also discovered that older persons trust automation more because of age-related cognitive deficiencies, which decrease their ability to detect automation flaws (Pak, Rovira, McLaughlin, and Baldwin 2017). Older adults welcomed automation and were more willing to off-load tasks to automation as it allowed them freedom in their work, regardless of the reliability of the automation (Soeters, Winslow, and Weibull 2006). For instance, studies on vehicle automation and decision aids indicate that older adults trust automation more as compared to younger adults (Donmez et al. 2006; Ho, Wheatley, and Scialfa 2005). However, these results were only limited to these specific contexts and may not be true for all cases.
- Gender, Ethnicity, Personality and Trust Propensity – These factors have been discussed elsewhere (Schaefer 2013; Schaefer et al. 2016). However, there are currently limited study on these factors in the automation/AI literature. These factors which are also beyond the scope of this thesis will be excluded from further analysis.

(2) States

The relationships between human states such as attentional control, fatigue and stress levels, and automation are comparatively well-researched areas of HMI. These factors all have an influence on the trust development between humans and the AI system.

- Attentional Control – Attentional control is defined as the ability of a person to pick what he pays attention (concentrates on) and what he ignores (Schaefer et al. 2016). Multiple studies indicated that human operators who exhibited lower attentional control interacted in a different way with an automated system. They

also relied more heavily on the automated system as compared to those with higher attentional control (Chen and Terrence 2009; Chen and Barnes 2012).

- Fatigue – Fatigue can be categorized into active and passive fatigue. Active fatigue is a stress-like state due to mental overload while passive fatigue is a result of underload and monotony. Existing research seems to suggest that fatigue leads to over-dependence on automation even though there are differing perspectives on the effects of active and passive fatigue (Matthews et al. 2019). Regrettably, the limited literature studies provide conflicting results. For instance, Reichenbach, Onnasch, and Manzey (2011) found out that contrary to what was suggested earlier, sleep-deprived (active fatigue) operators were more attentive towards possible automation errors and failures and showed less over-dependence. Operators who worked with a lack of sleep interacted more cautiously and intently with the decision aids and invested more time and attention to check the system COA and recommendations before approving it.

Over-trust and under-trust that contribute to over-reliance and under-dependence of the system may worsen performance (Lee and See 2004). When a human operator experiences high cognitive workloads, active fatigue may worsen over-trust issues, especially if the task requires the operator's vigilance. When automation was designed to remove some workload burden from the human operator, fatigued operators may become dependent on the system and susceptible to delegate excessive tasks responsibility to it, minimizing his cognitive demand. Hence, in situations that require a high cognitive workload, there is motivation to off-load some tasks functions to automation.

- Stress – In this thesis, stress is assumed to be induced by the operational tasks. Research has suggested that stress is dynamic and can lead to a decrement in trust (Schaefer and Scribner 2015). Biros, Daly, and Gunsch (2004) found that high task loads will lead to high stress and negative impact on human trust in

automation. However, it was also found out that during times of increased task loads, operators may resort to using system automation even though they have lower perceived trust in the system's automation. However, this is not universally true. In cases where the tasks are allocated effectively between humans and automation, an increase in task load will not degrade the trust and overall performance. Many times, the loss of SA due to inefficient automation monitoring results in performance decrements and a declining understanding of the work environment which decreases trust (Cosenzo et al. 2006).

(3) Cognitive Factors

The cognitive factors considered in this work include an operator's understandability of automation, and ability to use the system and expectation of the automation (Schaefer et al. 2016).

- Understanding how the automation work – The ease with which the human operators can learn to operate and engage with the automation determines how well they understand the automation. Past experience with similar automation systems also leads to understanding (Schaefer et al. 2016). Several studies show that trust develops over time, and previous encounters with similar technologies or systems have an influence on trust development (Merritt and Ilgen 2008; Rajaonah et al. 2008). However, the relationship between ease of learning and trust is not so clear. On one hand, Wang, Jamieson, and Hollands (2011) suggested that ease of learning was negatively correlated to trust. On the other hand, researchers have also found that operator's trust increases with training (Muir and Moray 1996). A study was done by Soeters, Winslow, and Weibull (2006) discovered that military cadets trusted information automation far more than decision automation. Decision automation becomes less explicit and more abstract as it replaces higher-order thinking skills like decision-making. The lack of transparency is a result of the complex computational algorithms and complicated sensor data fusion.

- Ability to use – The operator’s efficacy, workload, and competence all influence the “ability to use” of automated systems. Automation is frequently adopted to help humans to reduce their workload. It has been found that automation is effective in reducing both human physical and mental workload (Schaefer et al. 2016). In a study on combat identification tasks, Wang, Jamieson, and Hollands (2011) found that trust decreases if the automated systems result in an increase in cognitive workload. Moreover, the mental workload has a direct relationship with expertise with automation. Increased competence with an automated system, for example, often enhances trust development (Rajaonah, Anceaux, and Vienne 2006). However, more expertise may diminish a human’s capacity to watch for unexpected states in highly dependable automation systems (Bailey et al. 2006).
- Expectation – The expectancy of the automation system is influenced by a person’s mental model of the system, its perceived usefulness, expectations, its reputation, and the perceived advantages of the system. The perceived usefulness of an automation system is frequently judged based on its function. The perceived benefits of the automation system will affect whether it will be trusted (Lees and Lee 2007). Moreover, the reputation of an automated system may influence the human operator’s expectations of the system. Generally, there will be more trust when the system is known to have a positive reputation (de Vries and Midden 2008).

(4) Emotive Factors

Emotive factors considered in this work include attitudes toward automation, commitment to, and comfort with automation.

- Attitude toward automation – The attitude toward automation affects the overall awareness of system behavior. It was found that positive attitudes such as the happiness of an automation system affect trust, reliance, and over-reliance on the system (Bailey et al. 2006; Schaefer et al. 2016). On the contrary, errors or

difficult-to-use systems frequently give rise to negative attitudes (Ardern-Jones et al. 2009) and result in disuse (Gao, Lee, and Zhang 2006).

- Commitment – In the literature, commitment is often studied in terms of feeling satisfaction or delight. Research has shown that human satisfaction influences the human operator’s view of automation in many applications. For instance, satisfaction has been demonstrated to improve trust development with combat identification (Wang, Jamieson, and Hollands 2011) and vehicle automation (Donmez et al. 2006).
- Comfort – Comfort of using a system can influence a human’s perception of the system and can ultimately affect trust. It was found that the operator’s familiarity with and closeness to the system influence the level of the operator’s comfort with automation (Schaefer et al. 2016). However, there has not been a conclusive relationship between comfort and trust and more investigation is required.

b. Partner-Related Factors of Trust

The “partner” in this study refers to an automated system that works alongside the human as a teammate. Schaefer et al. (2016) and Hancock, Billings, and Schaefer (2011) hypothesized that the human-robot trust is moderated by the features and capabilities of the automation system.

(1) Features

The features of the AI system are important factors in the trust development between humans and the AI system. In this work, these features include mode of automation, appearance/anthropomorphism, and level of automation (LOA).

- Mode of Automation – The mode of automation includes fixed, adjustable, and adaptive. Generally, humans exhibit greater trust in AI systems when there is some degree of collaboration with the system. For instance, it was found that

humans generally have greater trust in AI systems with manually adjustable automation where humans have explicit control of the system, compared to implicit control (Moray, Inagaki, and Itoh 2000; Sauer, Nickel, and Wastell 2013).

- Appearance – The appearance of the automated system has been shown to influence automation appeal and use (Li and Yeh 2010). Pak et al. (2012) demonstrated that higher levels of human-like characteristics significantly enhanced trust development in complex systems. The physical form of the AI system should meet the human’s expectations required for specific functions (Schaefer et al. 2016).
- LOA – The LOA refers to the amount of control or level of autonomy the AI system is designed to exercise (Parasuraman, Sheridan, and Wickens 2000). In general, no direct relationship has been demonstrated between trust and LOA. However, it has been demonstrated that when humans perceive an automated system to be “good,” they are more likely to trust it than when they believe it to be “bad” (Merritt et al. 2013). Thus, it is crucial to appropriately design the LOA to perform the required function in helping the human operator to perform his work.

(2) Capabilities

One of the more significant influences on human-automation trust has been found to be an automated system’s capabilities. For instance, the consistency and reliability of an automated system over time is one of the most conclusive criteria that have been explored to increase trust in automation systems (Schaefer et al. 2016). Further investigation into the topic of system consistency and reliability reveals the importance of system feedback and error reduction in enhancing confidence. Three important factors of system behavior, reliability, and feedback that define a system’s capability are discussed in this section.

- Behavior – A system's behaviors should be predictable and dependable to positively influence trust development sustenance. A predictable system refers to a system that matches its operator's expectations. A dependable system refers to a system that is consistent and effective in performing the allocated functions. Operators will begin and continue to trust a system if the experience with the system provides predictable outcomes (Cahour and Forzy 2009; Muir 1994). However, there will be a drastic drop in trust which will result in disuse or disregard of the system when the operator experiences unanticipated reactions from the system (Schaefer et al. 2016). Hence, consistent and effective system behavior and actions are desired (Yamada and Kuchar 2006). Tenney, Rogers, and Pew (1998) observed that pilots prefer dependable automation that provides information that is both simple and informative.

For automated systems that provide alarms to users, the frequency and duration of the alarm affect trust development. In a study on collision avoidance systems, Gupta, Bisantz, and Singh (2002) observed that a reduced number of alarms resulted in higher trust. Disproportionately more alarms raised the levels of false alerts and were seen as unnecessary (Wickens et al. 2009).

- Reliability/Errors – Reliable systems have been shown to engender trust across different task contexts (Schaefer et al. 2016). Despite the fact that faults do occur in systems, the ability to reduce all types of errors boosts the system's reliability, which helps trust formation. (de Vries, Midden, and Bouwhuis 2003). On the other hand, it was discovered that a high miss rate and a high number of alarms erode trust and worsen performance (Wickens et al. 2009; Yamada and Kuchar 2006). Furthermore, a study has shown that the task difficulty level, rather than the type of error, appears to determine the level of trust. In an identification task scenario, Madhavan and Wiegmann (2007) found that trust decreases when automated systems fail or cause false alarms while identifying a target that the operator believes is easily identifiable. However,

trust develops when the automation finds a target that the operator believes is difficult to identify.

In a study conducted by Soeters, Winslow, and Weibull (2006) on trust and reliability, they found that trust in automated systems formed by military cadets and older adults is generally insensitive to the system's reliability. Trust for this demographic remained relatively high even though the AI reliability was perceived to be lower. This observation was supported by another study which also found trust to be insensitive to reliability for a group of civilian students performing military tasks (Rovira, McGarry, and Parasuraman 2007). The military cadets' lack of sensitivity to reliability could be explained by the fact that they treat AI as a teammate and are influenced by the military organizational culture, which requires authority and comrades to be trusted. (Pak, Rovira, McLaughlin, and Baldwin 2017).

- Feedback/Cueing – Research showed that the appropriateness of a system's feedback and cues directly affect trust development (Schaefer et al. 2016). Examples of this include the following:
 - Quantity of feedback (Muir and Moray 1996)
 - Information Accessibility (Bitan and Meyer 2007)
 - Type of feedback (Wiegmann et al. 2006)
 - Feedback accuracy and truthfulness (Sharples et al. 2007; Spain and Bliss 2008)
 - Effective communication to minimize conflicting information to operator (Stanton, Young, and Walker 2007; Song and Kuchar 2003)

c. Environment-Related Factors of Trust

In addition to factors relating to humans and AI systems, the operational environment that the system is operating in also affects trust development. Hancock et al. (2011) identified environment-related trust elements in terms of team collaboration and task/context.

(1) Team Collaboration

Team collaboration has become a highly researched topic in the realm of HMI as automated systems become more advanced with higher levels of autonomy. A review of the literature on team collaboration between humans and automation systems suggests the importance of cultural/societal impact and mental models.

- Cultural/Societal Impact – The society and culture that the operator is exposed to, impact trust development. The community, culture, health and well-being, environment, political system, and anxieties and ambitions about safety and the future all have societal consequences. (Clavell and Frowd 2014). A person's society creates a set of cultural norms that affects one's acceptance of an automated system (Huerta, Glandon, and Petrides 2012). Societal factors impact individual's mental models of automation.

An example is a military culture and its impact on trust development in automated systems (Soeters, Winslow, and Weibull 2006). The military culture is hierarchical, with strong social orders that adhere to laws and regulations while maintaining strict discipline. Military culture and the training that soldiers undergo can affect soldiers' expectations and behaviors with AI systems (Kennedy, Sibley, and Coyne 2015). Soldiers are likely to regard AI systems as comrades. Because of the AI system's teammate attitude, soldiers may initially have unreasonable expectations of the AI system's capability. (Soeters, Winslow, and Weibull 2006).

- Mental Models – The mental model of an automation system that an operator has can be influenced by societal impact, prior encounters with similar systems, cultural beliefs, and prejudices. System feedback can also be designed to update an operator's mental model of the system (Cuevas et al. 2007). Individual and collective mental models are structured pieces of knowledge that humans have that define and forecast the objective, form, function, or state of a system. (Rouse and Morris 1986). The mental model is discussed under environmental

factors as it is discussed at the society level instead of at the individual level. Mental models have been shown to have both direct and indirect influences on trust development. For instance, Fogg and Tseng (1999) found that if a design interface of a system matches the operator's mental model, the operator will perceive the system as credible and thus trust the system more. Also, Soeters, Winslow, and Weibull (2006) discovered that military cadets are more likely to dislike a system because of previous experiences with similar systems that resulted in terrible repercussions. (Pak, Rovira, McLaughlin, and Baldwin 2017). Other research also suggests an indirect relationship between the mental model and trust. Research on cockpit displays (Sarter, Woods, and Billings 1997) and driving automation (Walker, Stanton, and Young 2001) found that mental models are associated with reduced attention allocation of task elements. This will degrade the situation awareness of the operator which in turn decreases trust development in automation.

(2) Task/Context

According to human-automation trust research, risk and uncertainty are essential components in the formation of trust.

- Risk/Uncertainty – Risk can develop from both internal such as data corruption and external sources. Risk directly influences the evaluation of system safety which affects trust (Lin, Bekey, and Abney 2008).
- Context/Task type – The many task and function components of automation are critical in establishing trust in these systems. This section delves into three major areas of automation: cognitive assistance, control aids, and perceptual aids (Schaefer et al. 2016).
 - Cognitive aids strive to reduce the cognitive effort of humans by making recommendations concerning current and probable future conditions. During monitoring, these aids can perform accurate identifications or detections. Research has shown that transparency, reliability, operator's

experience, age, task difficulty, and risk have impacts on trust. For example, as previously noted, a reliable and dependable system fosters greater trust. Similarly, a transparent decision aid also contributes to higher trust (Mercado et al. 2016). In a study on using automated aid for inspection tasks, Madhavan and Wiegmann (2007) discovered that more experienced operators trust an automated aid more than less experienced operators. In addition, Ezer, Fisk, and Rogers (2008) found that older people tend to trust cognitive aids more than younger people.

- The goal of control aids is to replace certain degrees of human work, such as navigation and vehicle control. Trust in these aids is generally high and the default trust on these systems tends to be higher than a human aid (Ma and Kaber 2007). Like cognitive aids, the reliability of control aids was proven to have considerable influence on trust. Higher reliability has been shown to contribute to greater trust in fault management systems and patient monitoring systems (Moray, Inagaki, and Itoh 2000; Spain and Bliss 2008). In addition, older persons were found to have more faith in driver mitigation technologies than younger adults (Donmez et al. 2006). Additionally, this research also revealed that communication played an important role. Humans showed more trust when the systems provided accurate and reliable information and used human speech. Last, in the context of a vehicle control system, trust improved when the system reduced human workload and was perceived as useful (Donmez et al. 2006; Spain and Bliss 2008).
- Perceptual aids help operators by providing alerts/alarms or by assisting with pattern recognition. A study has shown that drivers demonstrated a higher level of trust in alarms with a low sensitivity rate (Gupta, Bisantz, and Singh 2002). Additionally, the alarms should be prompt to alert the operator (Abe and Richardson 2006). Reliability was found to

be an important factor for trust. Last, it was also found out that explainability and understandability of what and how the system is doing led to higher trust (Koustanaï et al. 2012; Gunning and Aha 2019). For instance, Gunning and Aha (2019) found that users prefer explanations, even if they lower performance.

2. Objective/Subjective Factors of Trust and Trustworthiness

Sheridan (2019) studied the factors of trust and trustworthiness of automated systems and categorized them into objective and subjective factors. In his earlier work, Sheridan (1988) emphasized the significance of trust in military Command and Control (C2) systems. He then proposed seven objective attributes as summarized in Table 7. Objective attributes are measurements that can be measured objectively, in a way that operational measures of efficacy or performance may be obtained to characterize the system's level of trust. The factors are categorized into two groups: the trustworthiness of the system and the trust attributes of the operator. Sheridan (2019) asserted that it is important to differentiate between these two groups of factors that are frequently misunderstood.

Table 7. Objective Factors of Trust/Trustworthiness. Adapted from Sheridan (1988).

S/N	Objective Factors	Description
Group I - System's Trustworthiness: Properties of the System		
1	Reliability	Low automation error rate
2	Usefulness	The automation's ability to do what is most important
3	Robustness	The automation's capability and flexibility in performing a wide variety of task
4	Understandability/ Explainability	Automation transparency in exposing the reasons and processes of what it does
5	Explication of intent	What automation will do next is communicated to the user.
Group II - User's Factors: Trust attributes of the operator		
6	Familiarity	Whether operator has prior experience with the system or similar system

7	Dependence	The extent to which the user would employ the automation system in comparison to other methods of performing the specified task
---	------------	---

A decade later, Sheridan (2019) presented a more comprehensive list of factors influencing trust development. Automation, without a doubt, has become more powerful and clever in terms of what it can do for us, as well as more varied and less predictable in many respects. It is becoming increasingly difficult, if not impossible, to comprehend how automation makes decisions and takes acts. Therefore, Sheridan (2019) asserted that as these systems become more intelligent or “human-like,” more subjective factors related to human characteristics and morality will be relevant and the knowledge of these factors will be useful. As such, a set of subjective factors is summarized in Table 8, has been included to form a more comprehensive list in his study of trust in automation.

Table 8. Subjective Factors of Trust/Trustworthiness. Adapted from Sheridan (2019).

S/N	Subjective Factors	Description
1	Care	The extent to which the system cares about its user after considering the work objectives, user’s ability, and restrictions in overseeing the automation. For example, it will consider the user’s response time, preferences, and programming faults, among other things.
2	Liberty	The automation is adaptable enough to allow the user to program in a variety of ways, as well as providing alternative displays and interaction/control options centred around user’s preferences. In addition, the system remains resilient even when the user commits procedural mistakes. It offers advises on how to repair problems and make the contact easier.
3	Fairness	The automation is consistent, and the operator will not be required to operate with greater speed, expertise, or programming skill than is expected. It will notify the operator if a request or instruction is beyond the automation’s comprehension or competence. It will not take measures that contradict the operator’s apparent objectives until safety is jeopardized. In the latter case, the system will provide the rationale for deviation.

S/N	Subjective Factors	Description
4	Loyalty	The automation will keep a record of its interactions with the operator to behave as desired by user. It will anticipate user needs as much as feasible and be prepared to act when called upon by the user.
5	Authority	The automation will carry out the specified actions and decisions based upon knowledge base of the allocated task. It will optimize its decisions and actions based on speed, accuracy, and resources while keeping the operator's goals in mind. Otherwise, it will operate with clear default aims.
6	Sanctity	The automation will present easily understood visual display and vocal messages to the user. The level of feedback desired by the user will be supplied. Such communication will be orderly and direct.

3. Denning's Six-User level (subjective) Software Quality Assessment

As automation systems become increasingly complex, intelligent, and “human-like,” researchers’ focus has switched to investigating subjective aspects, such as emotive components, to better understand the dynamics between humans and automation systems. For instance, Denning (2016) asserted that software quality evaluation has shifted from traditional objective yardsticks of correctness, reliability, and usability to user-level assessment. The objective measures gave an assessment of the software’s objective properties. However, as Shewhart (1931) noted that the quality of a product is ultimately a subjective assessment made by the users reacting to it. Many software today fall short of the objective international standard ISO 9216 but are still widely adopted and trusted upon (Denning 2016). Denning (2016) developed a six-user level software quality evaluation to address software quality from the user’s perspective by addressing fundamental questions such as “Does this software aid me in my work?” and “Can I rely on it?” The six user-level software quality assessment framework is summarized in Table 9. In Denning’s (2016) model, the highest value obtainable is four where users are satisfied with the software as that the software has indeed been useful to the users. Also, there is an element of human acceptance and a sense of “Care” from the system which were what Sheridan (2019) had argued for. Although Denning’s work was on a generic software system, the ideas could

also be applied to automation or autonomous systems (Hawkley, Mares, and Giammanco 2005).

Table 9. Denning's Six-User Level Software Quality Assessment. Adapted from Denning (2016).

Levels	Software Quality Assessments	Remark
4	Software delights	Users are satisfied with the software as it performs beyond expectations. It produces new and unexpected positive effects to the users' work. The users feel that the producer understands the users' environment and the software is designed specifically for the user to solve problems.
3	Software produces no negative consequences	After utilizing the program for a while, users have had no service disruptions. Users recognize that software has been carefully thought out since it is secure against vulnerabilities and offers functionality that the user did not request but will assist the user in avoiding unfavorable future problems.
2	Software fits environment	Users can use the software with ease, and it fits the context that the users are operating in. This means that users do not experience steep learning curve in adopting this software and the use of this software directly helps the users to get work done.
1	Software fulfils all basic promises	Users assess that the software meets the standard requirements. The objective ISO standard facilitate this level of assessment.
0	Some trust, begrudging use, cynical satisfaction	Users have enough trust in the software to utilize it willingly. Many pieces of software were published with bugs and security flaws, and many were only fixed after users complained.
-1	No trust	Users do not trust the software as there are software bugs, malware resulting in system crash.

Denning (2016) designed levels to illustrate the software's varying emphasis on user pleasure. The greatest level—"delight"—climbs to the top level because of the user-producer interaction. A satisfied user will state that the producer understood the user's work and environment to design a user-centric software. Also, the producer is available to assist in user's problems and to seize opportunities and care for the user.

E. TRUST REPAIR FACTORS IN AI SYSTEMS

This section reviews the literature on trust repair models and trust repair factors. Figure 11 summarizes the trust repair model proposed by de Visser, Pak and Shaw (2018) and Schaefer et al. (2017). It involves three main stages: (1) Relationship Act, (2) Relationship Regulation Act, and (3) Net Victim Effect.

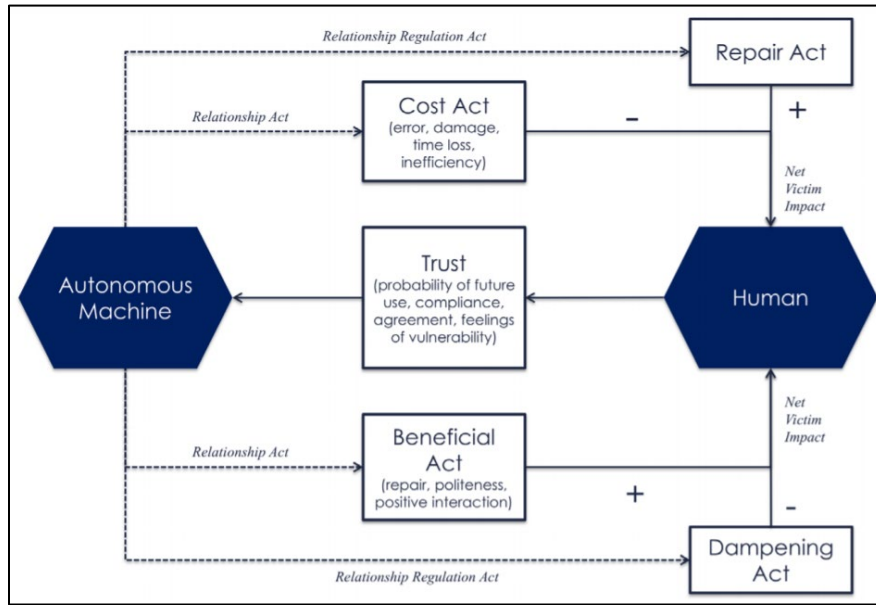


Figure 11. Trust Repair Model. Source: de Visser, Pak, and Shaw (2018).

1. Relationship Act

The cycle of trust repair starts with a relationship act performed by the autonomous system (de Visser, Pak and Shaw 2018). This relationship act can be beneficial or costly. Beneficial acts are those that people perceive as pleasant interactions. Examples of these include good, intended performance, politeness, and demonstration of algorithm capabilities. These positive activities can also serve as a general relationship mechanism, increasing credibility, goodwill, and forgiveness. On the other hand, costly acts are those acts that humans perceive to be harmful to trust. Examples of these acts are time loss, damage, inefficiency, miscommunications, and errors.

2. Relationship Regulation Act

A relationship regulation act is a corrective action taken immediately or after relationship activities (de Visser, Pak and Shaw 2018). The authors claimed that this regulation act is crucial for maintaining good and long-lasting relationships. Gottman and Levenson (1992) classified hypothetical regulating activities into two categories: repair and dampening acts. Repair activities reduce the negative effects of a costly act, whereas dampening measures reduce the impact of a favorable deed. Healthy partnerships necessitate these two types of activity.

3. Net Victim Effect

Researchers de Visser, Pak, and Shaw (2018) defined net victim effects as the combined effects of a costly and repair act or a beneficial and dampening act on human perception and experience. The authors also argued that each human will experience different impacts as individual differences affect trust. As a result, there is a need to incorporate this feature into the trust repair model because individual variances in readiness to resolve and rebuild the relationship may exist.

4. Repair Strategies

Table 10 provides an overview of the types and descriptions of trust repair studies conducted by different researchers. This table provides an example of many types of repair acts.

Table 10. Trust Repair Types and Descriptions. Source: De-Visser, Pak, and Shaw (2018).

Trust repair name	Description
Ignore	Machine deliberately ignores the occurrence of the costly act
Apologise	Machine conveys regret about the costly act and takes responsibility for its occurrence
Deny	Machine denies responsibility for the costly act
Empathise	Machine expresses empathy for the occurrence of the costly act
Emotionally Regulate	Machine identifies negative trigger and adds normative statement to stay positive
Recognize	Machine acknowledges that it performed a costly act
Blame	Machine outwardly blames human for the costly act
Anthropomorphise	Machine responds using a human communication channel without changing any other aspect of its behavior
Trump	Machine emphasises that the feature responsible for the costly act is in fact a strength, not a weakness
Explain	Machine provides explanation for why it failed
Promise	Machine makes a statement that it will do better in the future
Downgrade	Machine downplays significance of the act
Gaslight	Machine subtly suggests human should take responsibility for the costly act

It should be noted, however, that in the case of more complicated automated systems, basic notices or explanations may be insufficient to restore trust. As a result, when automated systems fail, deliberate actions must be taken to restore trust. The method must respond appropriately to the kind and severity of the trust violation and scenario. It is challenging for a system designer to find the appropriate trust repair activities due to the wide range of possible responses and dynamic parameters. It is yet too early to propose definitive solutions to this problem because research is still ongoing. de-Visser, Pak, and Shaw (2018) offered some general guidelines to automated system designers based on literature studies as follows:

First, there is no such thing as a one-size-fits-all trust repair technique that can be applied to every system. Because of variances in cognitive abilities, personality, experience, humans have the tendency to respond to apologies in a variety of ways.

Second, the trust restoration approach should be precisely matched to the type of trust violation by the autonomous system designer. de-Visser, Pak, and Shaw (2018), for

example, proposed that when an automated system commits a competency-based failure, such as a malfunction or output mistakes, any trust repair act is preferable to none.

Third, the context, which includes the surroundings and the precise situation of trust violation, should be investigated and comprehended. According to research, context influences how humans understand and respond to automation failure (Hoff and Bashir 2015a; Mittu et al. 2016). Furthermore, the degrees of risk associated with operations have a substantial impact on trust repair (Satterfield et al. 2017).

Fourth, the timing of a repair act is critical. According to a current study, apologizing at the next decision point for people is more effective than apologizing immediately after the trust violation in maintaining confidence. One possible reason is that humans do not remember the apology, but rather digest the information when it is important to their immediate action. The system will then require a model that detects trust violations precisely and alerts the user the next time a decision must be taken.

Finally, for effectiveness, automated system designers should examine the frequency of the trust restoration actions and the variability in communication (de-Visser, Pak, and Shaw (2018). Variability in how the trust-repair act is communicated or the source of the problem is revealed may result in greater success. Also, multiple apologies on the same problem without a change in system behaviors may reflect poorly on the system and make the system “insincere”—unwilling to change.

F. KEY TAKEAWAYS FROM CHAPTER II

This section summarizes the key points of Chapter II. These pointers will be useful for Chapter IV and V of this thesis.

(1) Summary from the Review of Trust Definitions

In the review of trust definitions, the concept of default trust is useful in the study of this thesis, as operators are likely to have been trained or operated with similar systems that may shape his/her impression of the AI-AMD system. As such, there is a need to investigate if default trust applies to the AI-AMD system in a military context.

Second, the definition of trust adopted in this thesis is “an attitude that AI-AMD system agent will help to achieve the operator’s goals in a situation characterized by uncertainty and vulnerability.” Trust is crucial in AMD operations as a high level of trust reduces uncertainty by anticipating the impacts of certain behaviors that make it easier for humans to work alongside AI-AMD systems in achieving missions.

Third, the concept of contractual trust states that trust exists if the AI-AMD system adheres to a specific contract. In this thesis, contracts refer to specific functions and behaviors that the AI-AMD is supposed to carry out.

Fourth, trust repair will be considered, which is defined as the act of improving trust after a trust violation has happened. Trust violations are inevitable as no system developed will ever be perfect. Hence, measures to address trust violations should be explored to maintain trust.

Fifth, it is noteworthy to point out that trust is different from being trustworthy. A system can be designed to be trustworthy but that does not guarantee that it will gain trust from its operator. However, a system must be trustworthy first for warranted trust to develop. Hence, this indicates the need to explore factors from two perspectives: the human operators and the system.

Last, in the review of trust calibration and trust specificity, the concepts of over-trust and distrust were discussed. Mismatches between humans’ trust and system capabilities will result in undesirable consequences. While it is important to develop human-AI trust, the concept of over-trust, which leads to over-reliance suggests that there must be mechanisms in place to keep trust at the desired level.

(2) Summary from the Review of HMI

The four stages of activities in describing an HMI would allow designers to identify factors that can lead to a relationship with trust between an operator and a future AI-AMD system. The system design considerations reviewed are also useful in identifying the factors that are relevant in improving the relationship between human operators and future AI-AMD systems.

The deep dive into the study of the “selecting an action” stage of HMI, refines the human information processing steps. Mapping these steps into system functions allows designers to describe the interactions between humans and AI-AMD systems with more granularity. Relevant trust factors that influence these human information processing step system functions can be identified to improve trust.

Pritchett and Feary’s (2011) work on the constructs of effective HMI provide valuable insights to improve the interactions between humans and system. For instance, the construct of representations is crucial, as the inputs and outputs of systems must be readable and understandable for humans to achieve effective interactions. These are building blocks for creating trusting relationships.

Another takeaway from this review is that interactions between humans and automated systems encompass a variety of tasks. This differentiation of tasks is referred to as function specificity by Lee and See (2004), in which humans interact with specific functions rather than the complete system. If a particular function was deemed untrustworthy, this could cause the human to distrust the entire system (Abbass 2019). As such, effective functional allocation to manage connections should be designed to enable trustworthy human-AI interaction.

(3) Summary from the Review of Trust Factors

On top of the traditional objective factors associated with systems, the extensive review of trust factors in the literature suggests the increasing importance of human-related (subjective) factors on trust. This indicates that trust is a psychological construct and that humans are ultimately accountable for the success or failure of operations. As a result, there is a need to sufficiently study and describe the HMI to identify the aspects that are relevant to human-AI trust.

Second, the review suggests that context, or the environment in which the human and system are operating, is crucial in determining the establishment of trust. As such, attention will be given to describe the environment where the HMI is taking place. Thereafter, with the environment being defined, the associated factors can be identified.

(4) Summary from the Review of Trust Repair Factors

In the trust repair model presented, it highlighted the importance of keeping the trust level at optimum level. This reinforces the need to prevent over-trust which leads to over-reliance and distrust which leads to disuse. Also, the potential trust repair strategy will be dynamic in nature which depends on the environment and situation.

III. LITERATURE REVIEW – OVERVIEW OF AI SYSTEMS FOR AIR AND MISSILE DEFENSE

This chapter presents the second part of the extensive literature review which analyzed the AMD threat scenarios and operations. Figure 12 shows the flow of Chapter III. First, this chapter reviewed the likely threats that an AI-AMD system would likely be contested with. With that, the threat scenarios from the RED force were postulated. Additionally, a possible threat from BLUE force to the successful AI-AMD operation was also considered to provide a holistic evaluation of trust in the latter part of the thesis.

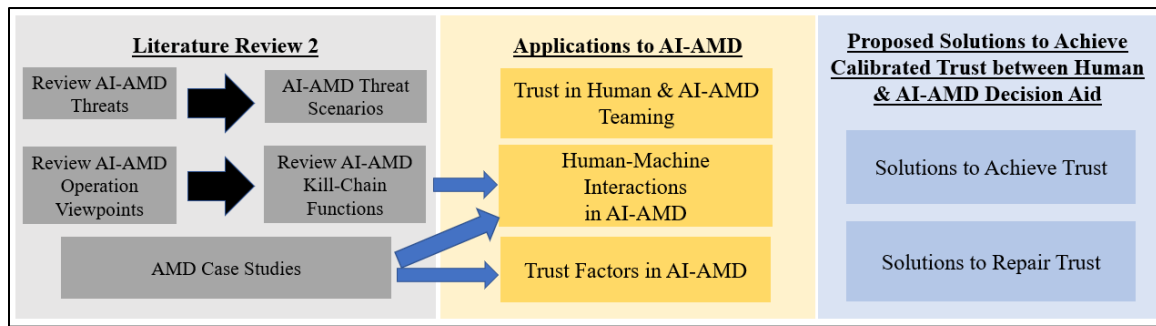


Figure 12. Schematic of the Flow of Chapter III.

Next, the DoDAF operation viewpoints of the AI-AMD were adopted from Jones et al. (2020) and presented in this chapter. The AI-AMD kill-chain functions were also reviewed to describe the HMI of an AI-AMD system. The chapter concludes with an AMD case study on the U.S. Army Patriot System to provide further insights on the HMI and trust of an AI-AMD system.

A. OVERVIEW OF AI-AMD THREAT SCENARIOS

1. Specific Threats

The U.S. military is expected to face wide-ranging threats from three main sources: (1) air threats, (2) ballistic missiles, and (3) rockets, artillery, and mortars (RAM) as prescribed in the Joint Publication (JP) 3–01 (Scott 2017).

- Air threats refer to manned and unmanned aircrafts and aerodynamic missiles.
- Ballistic missiles are those that do not generate lift from aerodynamic surfaces and follow a ballistic trajectory towards a predetermined target.
- RAMs are expedient, low-cost indirect fire means that have been typically employed against airbases, forward bases, and critical infrastructure.

Jones et al. (2020) had earlier in their study considered AMD threats from missiles and Unmanned Aerial Vehicles (UAV) for the application of AI on AMD systems. These threats and two other threats from manned aircrafts are compiled in Table 11. From Table 11, two observations are made. First, AMD threats have become more complex and harder to destroy due to their increased operational range and velocity. Second, air missile threats are far more demanding than aircraft and UAVs due to the high velocity which translates to short response time.

Table 11. Summary of AMD Threats. Adapted from Jones et al. (2020), Airforce-technology (2011) and Airforce-technology (2021).

Specific Threats	Operational Range (km)	Velocity (km/min)
DF-21D Hypersonic (China)	12,000	205.80
DF-17 Hypersonic (China)	2,500	205.80
WZ-8 Drone (China)	6,000	82.32
Chengdu J-20 (China)	3,400	35.00
P-270 Moskit (Russia)	250	61.75
BrahMos-11 (Russia)	600	144.07
3M-54 Kalibr (Russia)	50	16.08
Su-35 Multirole Fighter (Russia)	3,600	39.83
Khaliij Fars (Iran)	300	61.73
Shahed-129 Drone (Iran)	170	2.92

2. AMD Threat Scenarios

Two board categories of AMD threat scenarios from both the RED and BLUE perspectives were developed and discussed below. These threat scenarios indicate the need

for a high trust environment whereby the AMD operators could make decisions and respond to evolving threats promptly. Also, the methodology and training data used to develop the AI algorithm are important in building a trustworthy system in which the recommendations and actions are what operators would anticipate.

a. Traditional Threat Scenarios

First, traditional threat scenarios of missile threats of varying complexity towards the U.S. homeland were considered. These threats were further classified into three categories:

- Low complexity Scenarios – Single conventional missile threat such as non-hypersonic missile or fighter is the most common threat that the AMD encounters. The current AMD systems are equipped to deal with this threat as there is sufficient reaction time to destroy it because of lower velocity.
- Medium complexity scenarios – Multiple conventional missiles and/or fighters may pose a significant threat to the AMD system as the adversary may want to increase the likelihood of kill by launching multiple missiles simultaneously. These threats post mid-level complexity to the AMD operators. The mental workload of the operators increases with multiple targets as the operators must execute all the kill-chain functions on all targets simultaneously.
- High complexity scenarios – Advanced weaponry such as hypersonic missile is an emerging and complex threat to AMD. Hypersonic missile glides in the upper atmosphere and is highly maneuverable with a speed of at least five times the speed of sound (Mach 5). It can escape from the radar's detection during its flight phase and only remain detectable at the terminal phase of flight when homing towards its targets. As such, the operator is left with little reaction time to respond to such a threat.

b. Incongruent AI-AMD Decision Aid Recommendations from Operator's Concept of Operations (CONOPs) and Training, Tactics, and Procedures (TTPs)

In a scenario where a ballistic missile is launched towards the U.S. homeland. The operators and U.S. AMD systems are on alert supporting homeland defense operations. However, the AI-AMD decision aid recommends an engagement that is different from the operator's CONOPs and TTPs. This scenario creates a temporal trust deficit where the operator anticipates that the AI-AMD decision aid would propose a COA that is aligned to the operator would do. The operator will likely distrust the AI-AMD system and not adopt the recommendation. This scenario is analogous to the USS Vincennes incident where the Commanding Officer relied on his own judgement and took a different course of action.

c. Inadequate Trust Resulting in Catastrophic Consequences

The USS Vincennes incident highlighted that insufficient trust on the system outputs or recommendations may have dire consequences. The operators or commanders who ignore the system outputs may lose out on important information about the situation or environment. This may lower situational awareness. In situations where humans are unable to ascertain the information physically, this lack of trust and acceptance of system outputs may result in incorrect actions.

d. Over-Trust Resulting in Over-Reliance

On the other hand, operators who are over-reliant on the automated system may simply let the system run on automatic mode without checking its processes and outputs. The operator's skill will degrade after a long period of over-reliance on the system. The operator will not be able to detect any errors or incorrect system outputs. Since an automated system is seldom perfect and coupled with a dynamic environment, this may result in dire consequences without humans actively involved in the decision-making and executions process.

B. OVERVIEW OF AI-AMD OPERATIONS AND ARCHITECTURE

1. Operational Viewpoints and Scenarios

As illustrated in the previous sections, the fielding of more technologically advanced aircrafts and missiles has widened the scope and complexity of AMD operations. The U.S. forces must remain capable to counter these threats. The AMD system remains critical in achieving this mission to deny the enemy freedom of action (Scott 2017).

The AMD systems must then meet two key operational needs to counter these evolving threats. First, there is a need to shorten the operators' response time to respond to complex scenarios comprising of multiple and fast-moving targets. Second, there is a need to improve the operators' situational awareness (SA) by increasing the system surveillance range. To meet the requirement of improved SA, the AMD system is often part of a system-of-systems network where target information is shared amongst many systems. This allows the AMD operator to be able to "see" targets that are beyond its inherent detection range and prepare for tracking and engagement, if necessary.

An AI-AMD system would bridge the current capability gaps. An AI decision aid would allow for rapid information sharing across sensor platforms to increase SA. Within the AMD system, an in-built AI decision aid would reduce the operator's mental workload by providing recommendations that greatly reduces response time. In an earlier study by Jones et al. (2020), a model-based systems engineering (MBSE) architecture based on the DOD architecture framework (DoDAF) for an AI-AMD system was developed. The authors demonstrated that an increased level of automation of an AMD system significantly reduces the execution time to engage a target. The architecture frameworks developed by them will be reviewed in the following sections.

a. DoDAF OV-1: High Level Operational Concept

Figure 13 shows the high-level operational concept of an AI-AMD system. Boyd's concept of observe-orient-decide-act (OODA) loop was adopted to analyze the AMD operations (Blaha 2018). The OV-1 depicts that the AI-AMD system is part of a larger systems-of-system network of sensors, weapon platforms. The AI will enhance the operator's ability to "observe" and "orient" as information is shared across platforms and

organized for ease of understanding. Within the AMD system, the AI will also process this information quickly and automatically to provide the operators with a recommended course of actions to “decide” and “act” fast. As such, the OODA loop of AMD operations will be shortened with AI.

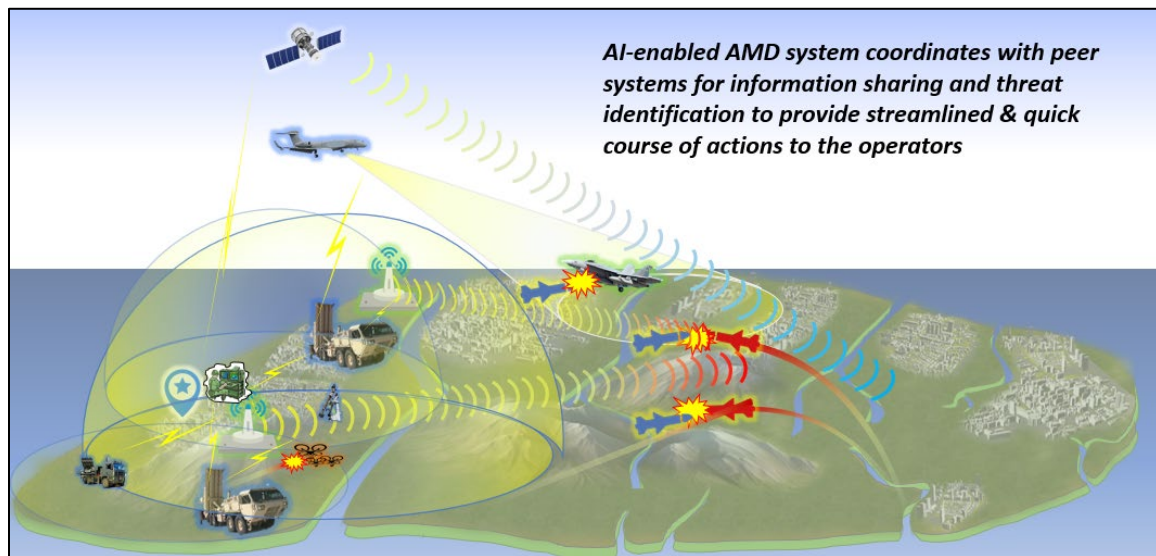


Figure 13. OV-1 of AMD Operations

b. DoDAF OV-6c: Operational Scenarios

Jones et al. (2020) developed an OV-6c (Event Trace Description) to illustrate the AMD operational scenario as depicted in OV-1. Figure 14 gives the OV-6c operational scenario that is based on JP 3-60 Joint Targeting. The OV-6c depicts the agents (AI-AMD decision aid, blue forces, red threat, sensors, and network) and their actions in an AMD scenario. Jones et al. (2020) selected a generic AMD threat scenario of a single AMD threat (hostile aircraft, UAV, or missile) and defined the sequence of actions that each agent undertakes in a typical AMD operation. The scenario is intended to be generic enough to include the planned sequence of events of other threat scenarios.

A typical AMD operation is as follows:

- It begins with a red threat launching an attack against BLUEFORCE (BLUFOR) assets.
- BLUFOR are unaware of the attack until the BLUFOR sensors detect it.
- Sensors send the detection data to the AI-AMD decision aid through a network.
- AI-AMD decision aid requests for more information from the BLUFOR sensors through the network.
- Sensors send their updated detection information back to the AI-AMD decision aid.
- AI-AMD confirms a valid target detection and requests the BLUFOR sensors to track a target.
- AI-AMD determines available BLUFOR assets for engagement and checks that there are no friendly forces within the area of operations and requests status and location of known forces.
- Once the AI-AMD de-conflicts the area, it issues orders through the network along with fire command to the BLUFOR attack mechanism.
- BLUFOR engage the threat, followed by an assessment of the engagement.
- If the red threat has not been destroyed, the AI-AMD can re-issue an attack command, and the BLUFOR engage again.

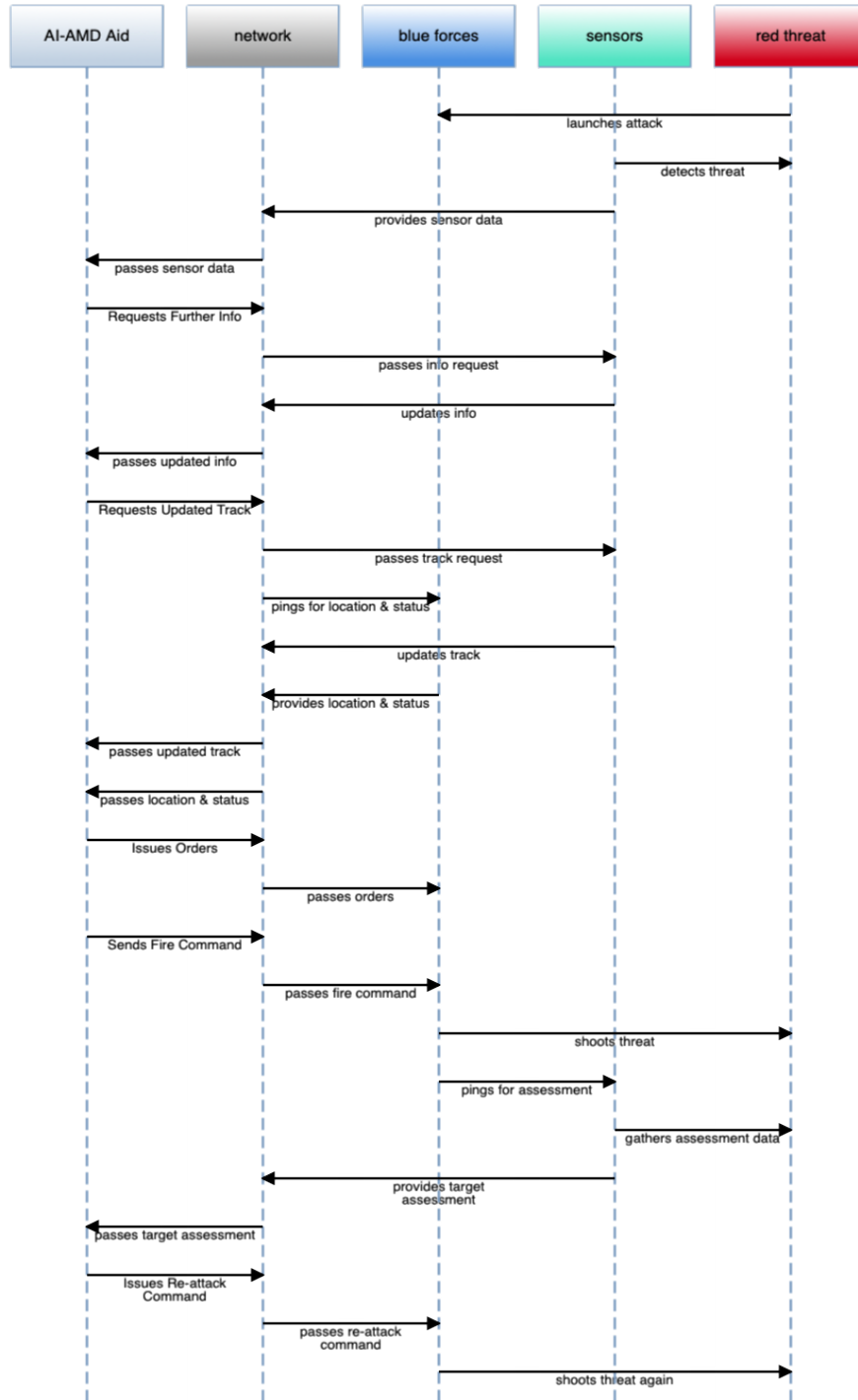


Figure 14. OV-6c: AI-AMD Event Trace for AMD Operations. Source: Jones et al. (2020).

c. DoDAF OV-5a: Operational Activity Decomposition Tree

Jones et al. (2020) identified 17 critical functions for the AI-enabled AMD system as outlined in the JP 3-60 process. Jones et al. (2020) streamlined these functions of the AI-enabled AMD system to higher-level functions from the JP 3-60 find, fix, track, target, engage, assess (F2T2EA) kill-chain as shown in Table 12. The F2T2EA processes were further streamlined to the OODA loop. Jones et al. (2020) further developed OV-5a, an activity decomposition tree that depicts the operational scenario from the OV-6c in a hierarchical function diagram. Figure 15 shows the internal capabilities decomposition of the AI-AMD system and Figure 16 shows the external system decomposition. The OV-5a depicted in Figure 16 defines the relationships and interactions the AI-AMD has with the external agents within the System-of-Systems (SoS) (i.e., BLUFOR sensors, weapons, and network) as well as the REDFOR threat.

Table 12. Mapping of Critical Functions to F2T2EA and OODA Loop

OODA	F2T2EA	Functions
Observe	Find	Collect Data
		Accept Initial Detection
		Identify Emerging Target
	Fix	Request Further Information
		Classify Target
		Locate Target
		Validate Detection
Orient	Track	Requested Updated Target Track
		Validate Target
		Assess Blue Proximity
Decide	Target	Nominate Engagement Options
		Prioritize Targets
		Select Attack Option(s)
Act	Engage	Issue Orders
		Send Fire Command
	Assess	Assess Target Status
		Authorize Re-attack

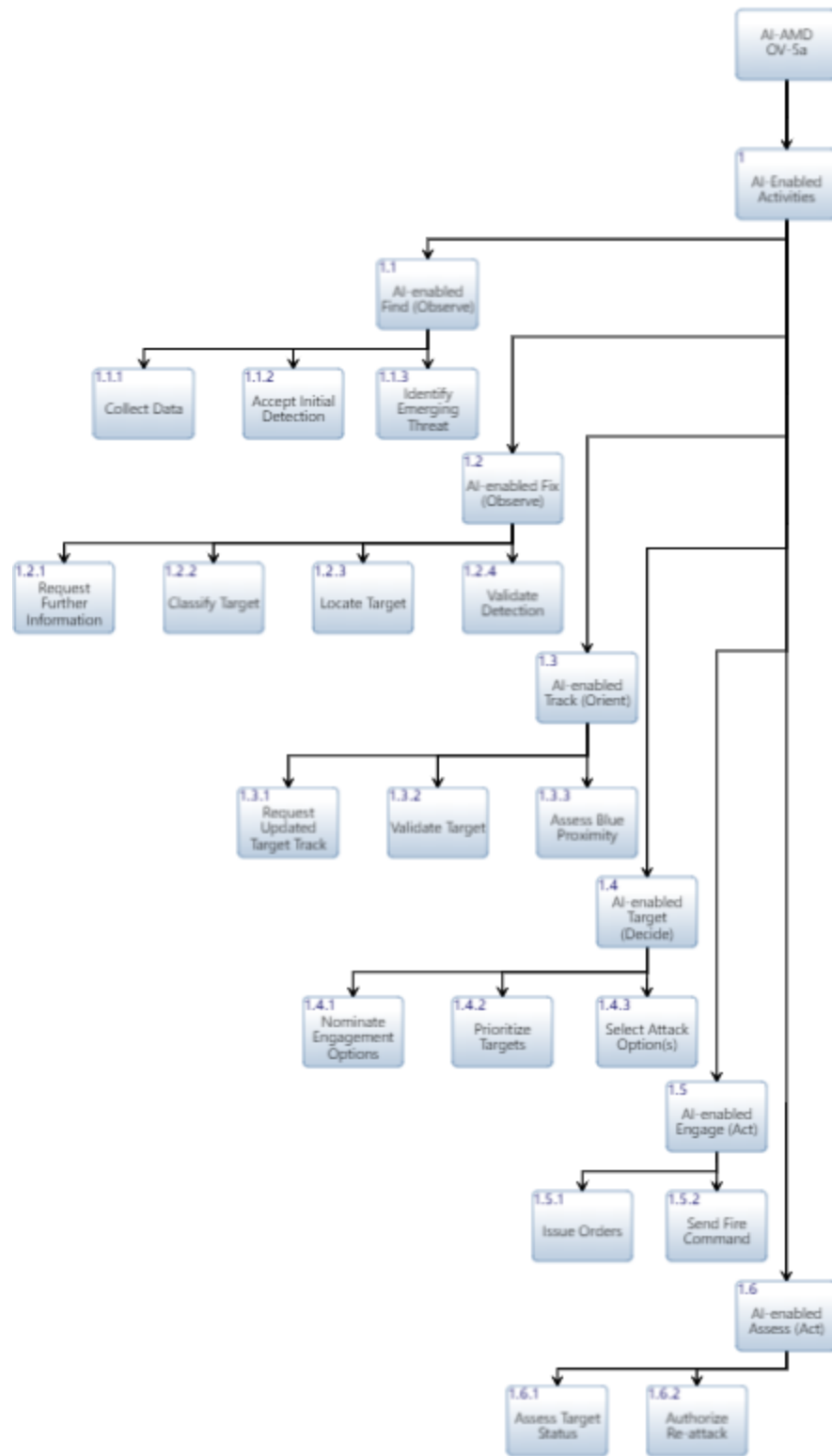


Figure 15. OV-5a: AI-AMD Operational Activity Decomposition - Internal Capabilities. Source: Jones et al. (2020).

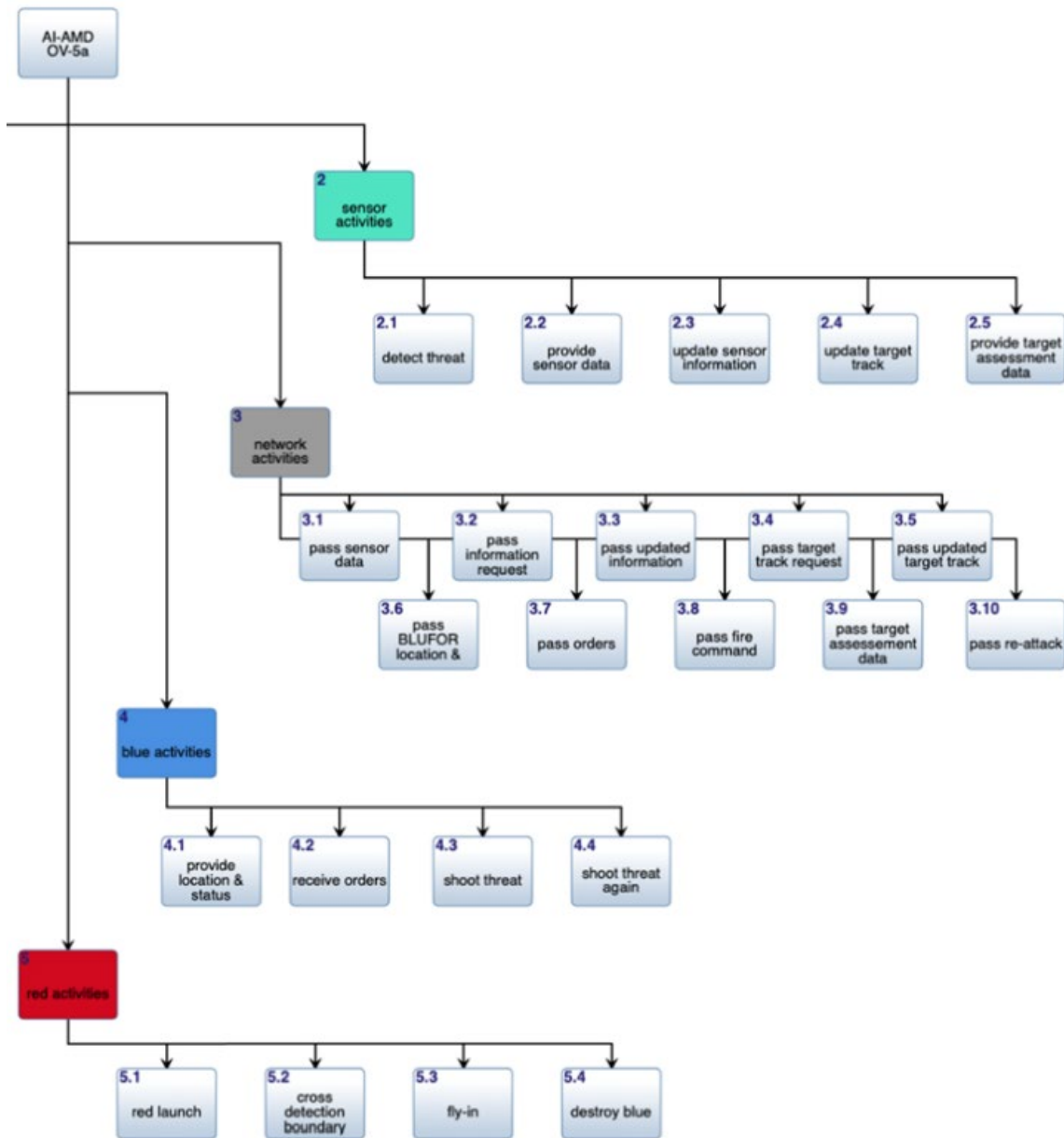


Figure 16. OV-5a: AI-AMD Operational Activity Decomposition – External Systems. Source: Jones et al. (2020).

d. DoDAF OV-5b: Operational Activity Model

The OV-5b activity model is built upon the relationship defined in the OV-5a and the inputs-controls-outputs-mechanisms (ICOM) between each agent. Figure 17 shows the high-level activities performed by the AI-AMD decision aid and external agents. Figure 17 also defines the inputs and outputs of each action performed by the agent, controls that trigger or limit the functionality, and the mechanisms needed to enable a particular function. It also demonstrates how the AI-AMD system communicates with other peer SoS systems via the shared network. Jones et al. (2020) then carried out a detailed study of the AI-AMD decision aid operational activities and its relationships with other peer SoS systems as depicted in Figure 18.

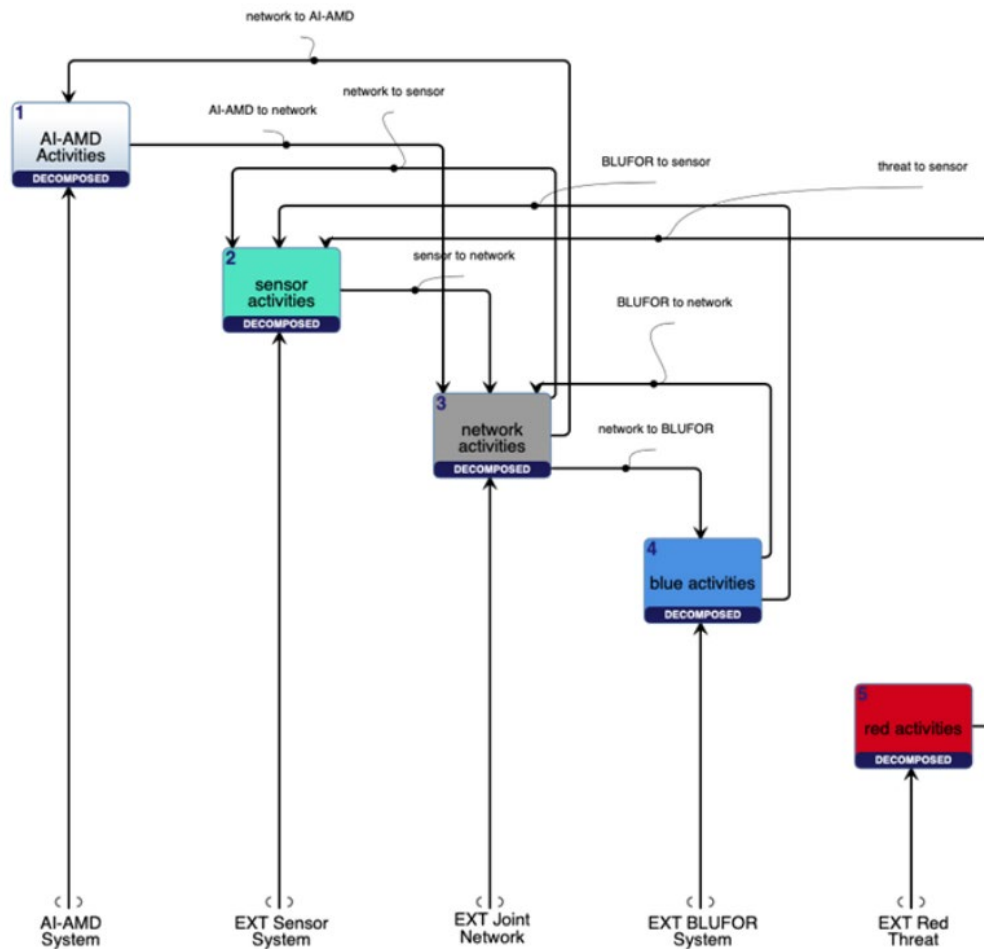


Figure 17. OV-5b: AI-AMD Operational Model. Source: Jones et al. (2020).

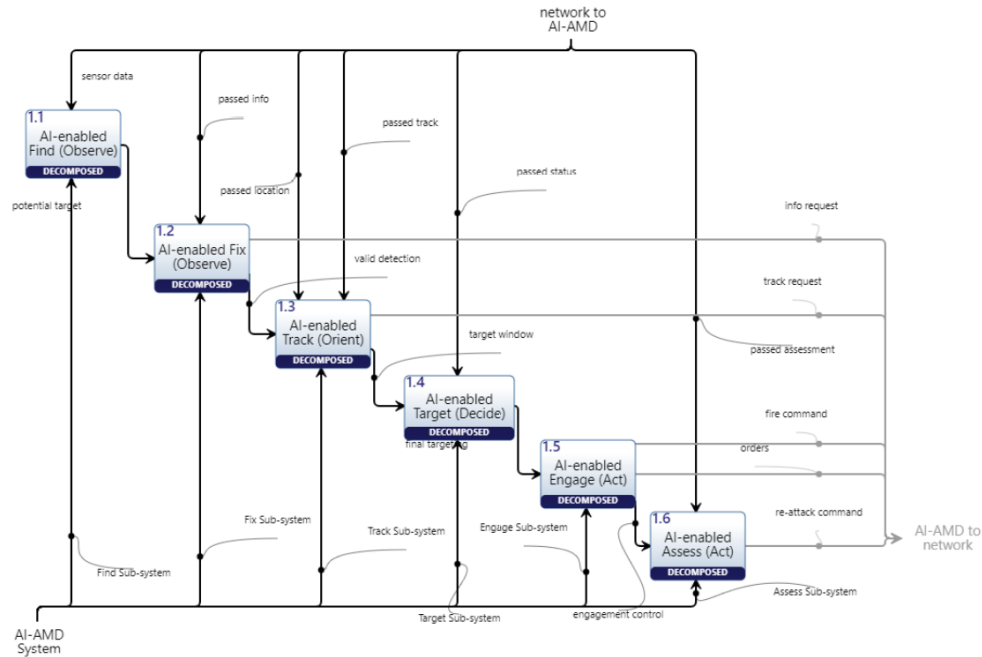


Figure 18. OV-5b: Decomposed AI-AMD Operational Activity Model.
Source: Jones et al. (2020).

e. DoDAF OV-5b/6c: Action Diagram

Jones et al. (2020) developed the OV-5b/6c by combining the operational activity model and event trace diagram to demonstrate the inputs and outputs of the operational activities and the sequence of events to establish the relationships amongst the peer systems. Figure 19 shows the set of AI-AMD top-level activities performed with the external agents. Green boxes represent the triggers such as exchanges of information or controls for enabling actions. Gray boxes represent the outputs of AI-AMD decision aid activities pass as inputs (gray boxes) into the next F2T2EA step.

The AMD operation commences when a REDFOR launches an attack against the BLUFOR. BLUFOR will be unaware of the threat until it crosses a detection boundary. Once the threat crosses the boundary, a “signature” trigger will be sent through the network to the external sensor system. The network will then receive raw data and pass it further to the AI-AMD decision aid. The AI-AMD will then request more information of the target to the sensor through the network.

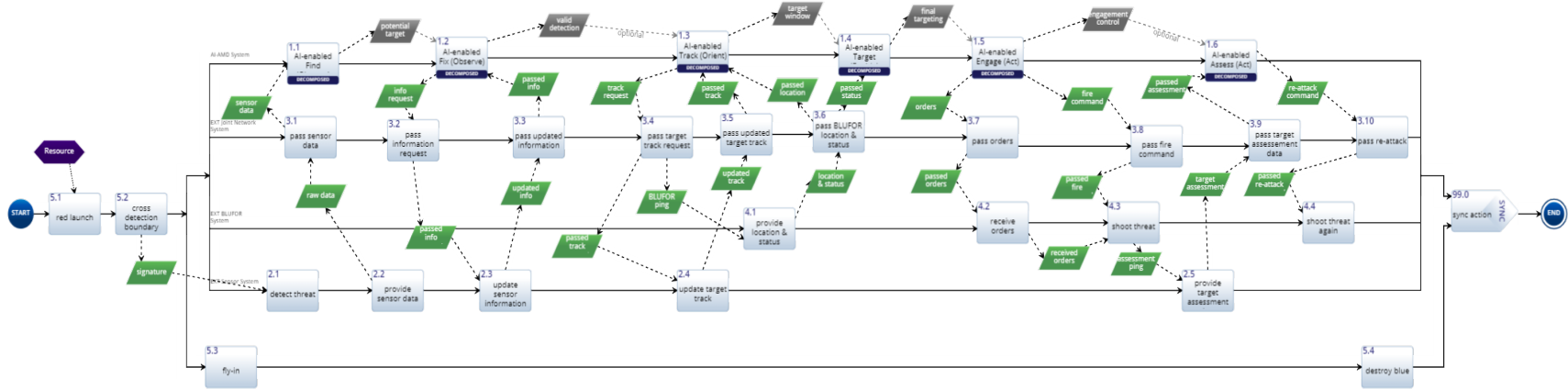


Figure 19. OV-5b/6c: AI-AMD Action Diagram. Source: Jones et al. (2020).

The network will then pass the sensor data to the AI-AMD decision aid with a “passed info” trigger. A “valid detection” output to track the target will be made after the AI-AMD decision aid successfully identify the target. The fix step is a continuous loop process where the AI-AMD system is continuously updating the decision aid based on sensor data.

The AI-AMD track step performs two key functions: (1) to update sensor detection tracks and (2) to determine BLUFOR defeat capabilities based on the target track. Once the track step is initiated, the AI-AMD decision aid will then send a “track request” to the network. The network will further send a “passed track” to the sensor to update the target track and a “BLUFOR ping” to BLUFOR assets to determine the location and status of BLUFOR assets. The sensor will then send an “updated track” to the network, and the network will send a “passed track” to the AI-AMD system. Location data from the external BLUFOR systems using the trigger “passed location” is used by the track step to update the “target window.” The track and target steps are in a continuous loop where information of the sensor detection tracks and BLUFOR location data are continuously being updated. Once the AI-AMD has established both the target track and BLUFOR location data, it will output the “target window” to the next step, target.

The AI-AMD decision aid will receive the BLUFOR location and status from the network using the trigger “passed status.” This information is obtained from the BLUFOR assets using the trigger “location and status.” The AI-AMD system will then generate Courses-of-action (COA) using the information obtained. It will also select an engagement and send a “final targeting” trigger to the next step, engage.

The engage step has two main functions: (1) to issue orders and (2) to send fire commands to the BLUFOR systems. The AI-AMD system will issue “orders” and “fire command” triggers to the network. The orders will then be sent to the BLUFOR system using the trigger “passed orders.” Also, once the fire command is received by the BLUFOR system using the trigger “passed fire command,” the BLUFOR system will shoot at the threat. The AI-AMD decision aid will then output “engagement control” to the next step, assess.

f. Decomposed F2T2EA Kill-Chain Functions

Jones et al. (2016) reviewed the JP 3-60 targeting doctrine and further decomposed each of the F2T2EA AI-enabled decision aid operational activities into lower-level activities. Figure 20 shows the decomposition of the F2T2EA Find step into three serialized activities; (1) Collect Data, (2) Accept Initial Detection, and (3) Identity Emerging Threat. The “collect data” activity is triggered by the “sensor data” input and concludes with an output of “potential target” which is a trigger to the Fix step.

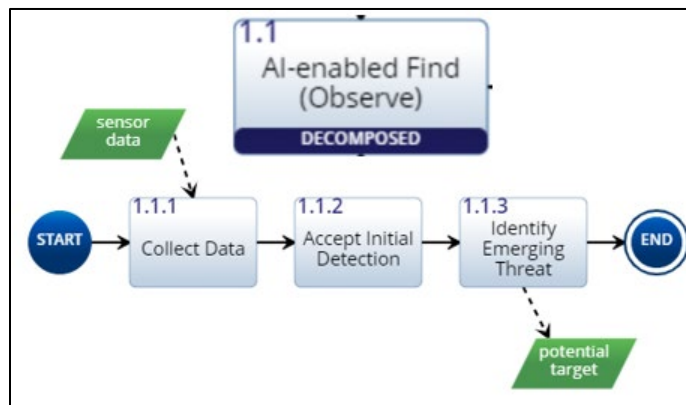


Figure 20. Decomposition of F2T2EA Find Step. Source: Jones et al. (2020).

Figure 21 shows the decomposition of the F2T2EA Fix step into four activities; (1) request further information, (2) classify target, (3) locate target, and (4) validate detection with (2) and (3) carried out in parallel. The Fix step has several triggers such as accepting the “potential target” from the Find step, “passed info” from the sensors to classify and locate target. This step concludes with a valid detection which is an input to the Track Step.

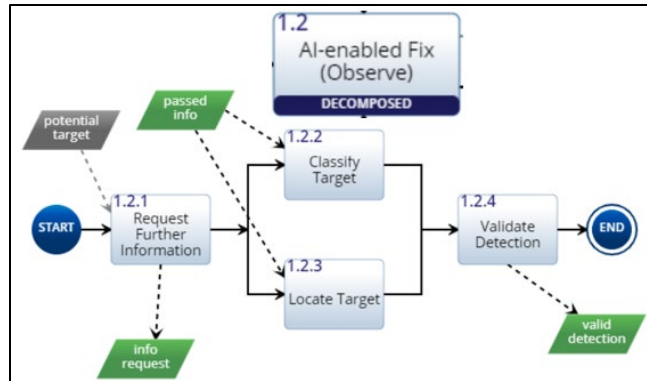


Figure 21. Decomposition of F2T2EA Fix Step. Source: Jones et al. (2020).

Figure 22 shows the decomposition of the F2T2EA Track step into three activities; (1) request updated target track, (2) validate target, and (3) assess blue proximity with (2) and (3) conducted in parallel. This step starts with a trigger of “valid detection” from the Fix step. A track request will be output to the sensors. The “passed track” trigger is used to validate the target and the “passed location” is used to assess BLUFOR proximity. The AI-AMD decision aid uses the BLUFOR data on location and the valid target to determine the “target window” to destroy a threat target.

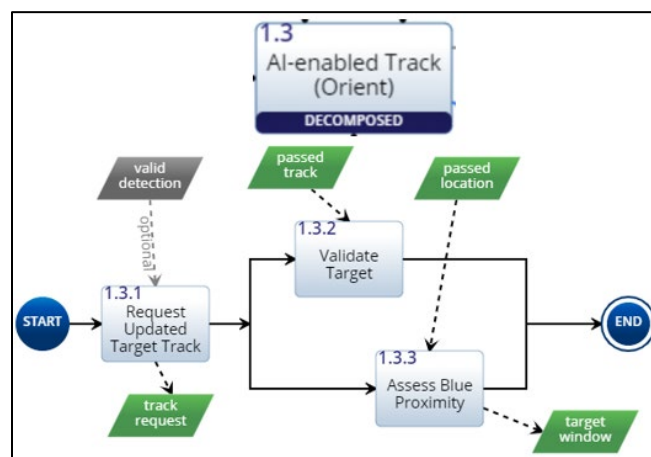


Figure 22. Decomposition of F2T2EA Track Step. Source: Jones et al. (2020).

Figure 23 shows the decomposition of the F2T2EA Target step into a series of three activities; (1) nominate engagement options, (2) prioritize targets, and (3) select attack options(s). The Target step starts with a trigger “target window” from the Track step. The trigger “passed status” will star the nominate engagement options activity. This Target step will conclude with a “final targeting” output passed to Engage step.

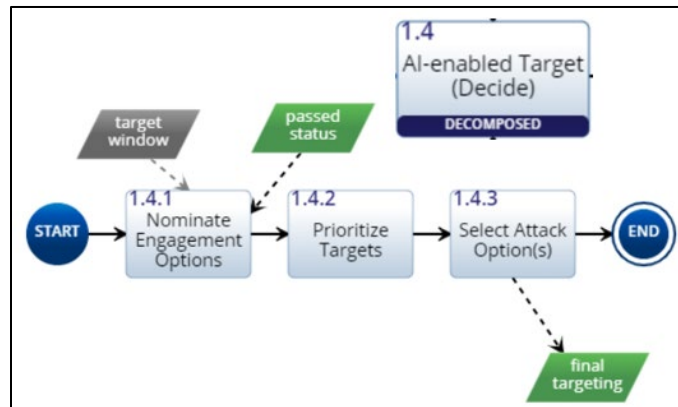


Figure 23. Decomposition of F2T2EA Target Step. Source: Jones et al. (2020).

Figure 24 shows the decomposition of the F2T2EA Engage step into a series of two activities; (1) issue orders, and (2) send fire command. This step starts with the trigger “final targeting” from the Target step to issue orders to the BLUFOR weapon systems. The AI-AMD will then output “fire command” to the BLUFOR weapon systems to carry out the engagement. This step concludes with an “engagement control” to the last step, Assess.

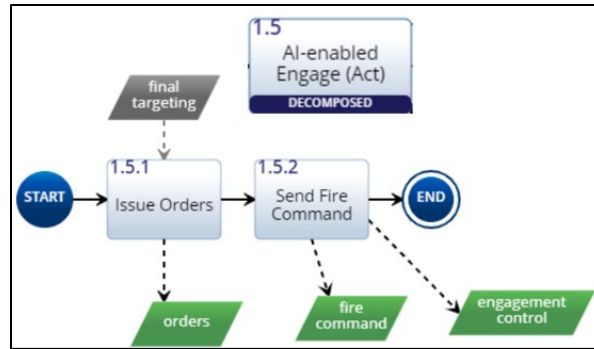


Figure 24. Decomposition of F2T2EA Engage Step. Source: Jones et al. (2020).

Figure 25 shows the decomposition of the F2T2EA Assess step into a series of two activities; (1) assess target status, and (2) authorize re-attack. This step starts with a trigger engagement control from the Engage step. The “passed assessment” trigger will then start the assess target status. If the REDFOR threat is not neutralized, the AI-AMD decision aid will output a re-attack command to the BLUFOR weapon system to re-engagement the target. This step concludes the entire AI-AMD operational activities within the OV-5b model.

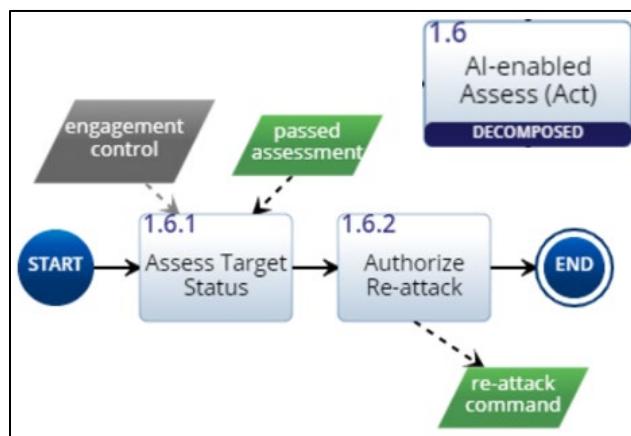


Figure 25. Decomposition of F2T2EA Engage Step. Source: Jones et al. (2020).

2. Characteristics of AI-AMD System

This section discusses several characteristics of AMD operations that are important to the study of human-AI trust.

a. AI-AMD Operations are Inherently Complex & Time-Sensitive

The AMD operations are inherently complex due to the element of surprise and short decision space. The AMD operations are highly reactive in nature as the BLUFOR are in a defensive stance. The fielding of more advanced weapons such as hypersonic missiles presents immense challenges to the AMD operator as the window of engagement is reduced significantly. Additionally, the operators may encounter multiple threats that require the operators' simultaneous response to each of these threats. The operator faces immense stress due to the information overload and increased mental workload in such a time-stressed scenario. As such, as the AMD operations become more complex, a well-established AI-AMD decision aid is required to help alleviate some of the operator workloads. Some tasks and decisions that were traditionally performed by the operator could be carried out or recommended by AI. These will help to reduce the time for decision-making. However, to produce a high-performance AI-AMD decision aid, there must be trust between the operator and AI-AMD decision aid.

b. Fully Autonomous AMD System will not be Viable in near Future

The AMD operations involve a function to destroy a target that takes human lives. This raises ethical concerns when an autonomous system is authorized to make decisions to kill humans. There had been instances of AMD fratricide such as the USS Vincennes incident and two U.S. Army's Patriot system fratricides on British Tornado and Navy F-18 during Operation Iraqi Freedom and Second Gulf War. The Patriot incident happened when the Patriot was placed on automatic mode (Hawley 2017). These incidents derail any efforts for the automation community to advance and field fully autonomous weapons. For instance, there had been no lethal autonomous weapon systems developed since the 1980s in the U.S. The argument against the use of these autonomous weapons is the perceived lack of accountability and a perceived inability to comply with international humanitarian laws. Also, as Scharre (2018) noted, the use of these autonomous weapons involves high

operational risks that could arise from software errors, unexpected environmental influences, and malicious enemy actions. As such, the human operator is still needed to operate in or on-the-loop of the AMD operations.

c. Human-AI Integration is Paramount for Successful AMD Operations

Since the AI-AMD operators operate the AI decision aid with human-like intelligence, the concept of an effective team would apply. The AI decision aid can be seen as the human teammate operating alongside the operators. The AI-AMD decision aid must be able to perform to what the operator expected and help improve the performance of the operator. Trust which is the foundation of an effective and successful team would thus apply to a human-AI team. The operator has many interactions with the AI-AMD decision aid through the functions that the operator must perform. The AI-AMD decision aid should be leveraged to perform the overall performance. The operator must also be clear on his involvement and decision authority in every phase of the operations. For instance, the operator must be clear when he is “in-the-loop” (the AI-AMD must wait for the operator to take any action before continuing) or “on-the-loop” (operator supervises the AI-AMD and intervenes when necessary) of the decision-making process to better appreciate the value of the AI-AMD decision aid and its responsibility.

C. REVIEW OF U.S. ARMY PATRIOT MISSILE SYSTEM

The U.S. Army Patriot system was one of the first lethal automated tactical systems to be employed in combat. The U.S. Army Patriot operates with two modes: automatic and semi-automatic. The semi-automatic mode is a human-in-the-loop system. The Patriot engagement-control algorithms and methods employed in the system’s automatic mode were modified from Safeguard, the United States’ first operational anti-ballistic missile system. Safeguard’s level of automation was not suited for Patriot as the mission and operating environment of the two systems were different. The Patriot system decision cycle is illustrated in Figure 26 and is similar to what was reviewed earlier in the chapter.

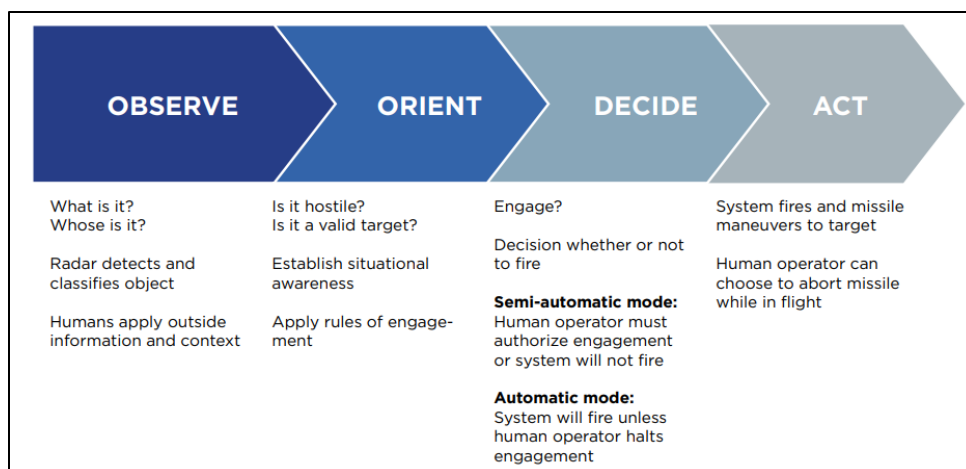


Figure 26. Patriot Missile System Decision Cycle. Source: Hawley (2017).

The Patriot was first used operationally during Operation Desert Storm against Iraqi Scuds (crude tactical ballistic missiles) in the early 1990s. The ballistic missile engagement scenario during Operation Desert Storm mandated more automation support. As a result, the Patriot was placed in automatic mode to illustrate the potential for an appropriate and timely new system. During Operation Desert Storm, the Patriot system successfully engaged scud missiles. However, the track classification and identification automatic mode had minor problems. These issues were overlooked since nothing disastrous occurred. Complacency set developed, resulting in an inappropriate trust and dependence on Patriot's automatic mode.

During Operation Iraqi Freedom, the U.S. Patriot was involved in two fratricide incidents, one with a British Tornado and the other involving a Navy F-18. After a post-incident review, it was found out that the U.S. Army had failed to adapt to a new complex capability of the Patriot's automatic mode. Specifically, the U.S. Army over-trusted the Patriot system and became over-reliance on it. They were not sufficiently prepared and trained for proper monitoring of the Patriot's automated operating mode and were unwilling or unable to distinguish on-the-loop and in-the-loop control (Hawley 2017).

This section discusses six key insights from this Patriot case study.

(1) Automated Systems Seldom Meet the Performance Requirements

Newly developed automated systems seldom live up to the expectation (Hawley 2017). Users need to decide how they should practically use the automated system rather than wholeheartedly using the system the way the developer thought it should be used. Hawley (2017) claimed that system developers are frequently unable to predict all the operational issues that an automated system may cause. “Surprises” from automation are to be expected. Accepting the system without question may result in automation misuse or over-reliance.

(2) Automation Usually Increases System Monitoring Workload

Automation may not always simplify the operator’s work as the nature of the operator’s job usually changes with automation (Hawley 2017). Automation is often associated with a complex system with many components. Many times, automation reduces the amount of time an operator spent on physically working on tasks. However, the operator must now monitor more indicators from the system due to an increased number of components. Prolonged vigilance on the system involves tough mental work and can be stressful. The author argued that it is cognitively challenging for the operator to sustain a high level of vigilance for an extended period when nothing out of the ordinary is happening. As such, it is unrealistic to expect a high level of vigilance from an operator monitoring the automated system for 23 hours 59 minutes and expect a high level of performance during that one minute of panic when a target is impending.

(3) Accept the Notion that Automation System can Fail

The belief that the system is always right will give the operator a false impression of the system. Hawley (2017) claimed that this will result in complacency and not conducting the necessary checks. A prolonged period of complacency will result in reduced vigilance and operational caution. Trust in the system must be built incrementally from experience and interactions and will always be context specific. He suggested that

operators must operate the system long enough to gain sufficient experience to determine when to trust the system and when to scrutinize and supervise the system closely.

(4) Automation Changes the Job Scope of Operator from Manual Control to Supervisory Control

The operator's job shifts from direct manual management of system operations to higher-level supervisory control because of automation. The operator is usually tasked to "monitor" the automated system if it is functioning and performing as expected and intervene when necessary. So far, humans have difficulty meeting the monitoring and intervention demands required by this on-the-loop control (Hawley 2017).

(5) Out-of-the-loop Familiarity

The operators' interactions and familiarity with the system will reduce when tasks are being allocated to automation. There is evidence that the operator will be slower to react and take appropriate actions during an abnormal situation (Hawley 2017). The operator loses SA, and the operator must scramble "back into the loop" to establish SA to make an appropriate decision. In the long run, operators may lose fundamental skillsets and proficiency in operating and controlling the system as they get less hands-on experience.

(6) Need for Increased Training Requirement

It is challenging for humans to meet the performance demands required for the on-the-loop supervisory control. The need for sustained vigilance and maintaining broad SA for interventions is a problematic challenge. It is unreasonable to expect humans to have reliable control over a complex automated system. As such, operator's training must adapt to meet the new challenges with automation.

An automation system that is usually more complex would cause operational skills to be more difficult to learn and retain. Also, the operator is required to have deep knowledge of the automated system he is operating for appropriate intervention when necessary. They must understand the "nuts and bolts" of how the system works. The current practice of frequent simulator training or in-the-loop training to deal with out-of-the loop

control skill decay is problematic. There are difficulties generating invaluable stressful scenarios which are essential for the operator to develop and maintain the necessary skillset for out-of-the loop control. Long-term training employing a hands-on instructional regimen that encourages operators to cope with difficult scenarios to develop their proficiency and understanding of the system can lead to expertise.

D. KEY TAKEAWAYS FROM CHAPTER III

This section of Chapter III summarizes the key points of this chapter. These pointers will be useful for Chapters IV and V of this thesis.

The review of the AMD threats and operational scenarios reveals the importance of AI to augment the operator to deal with increasingly complex situations. A key success factor for the operator to exploit the benefits of AI being applied to the AMD system would be a trusting relationship with a trustworthy AI-AMD system. Additionally, the inclusion of AI in the system will increase the complexity of the interactions between humans and the system. As such, a deliberate study of the HMI is necessary for effective interactions which is an ingredient for a trusting relationship.

The review of the operational viewpoints of the AMD operations led to the functional analysis of the AI-AMD system. It suggests that the AI-AMD C2 system through the network interfaces is the most critical aspect of the entire system as it coordinates the entire operations from detecting a target to engaging it. It produced 17-step kill-chain functions which will be used to describe the HMI between the operator and AI-AMD decision aid.

An in-depth study of the AMD operational viewpoints and the current acceptance of automated systems within DOD also highlights that a fully automated AI-AMD system will not be viable soon. Human is still required to be part of the system. As such, further analysis of the HMI between humans and AI-AMD across the 17-step kill-chain functions will have to be carried out. A clear level of human involvement and/or automation across all functions must be specified. Once the HMI is described sufficiently, further study of the relevant trust factors will then be identified to develop the strategy.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS OF TRUST, HUMAN-MACHINE INTERACTIONS (HMI) AND TRUST FACTORS OF AI-AMD SYSTEM

This chapter discusses the boundaries and environment that the AMD operations take place in which will bound the scope of subsequent analysis. Next, it presents a framework of trust associated with human and AI-AMD teaming. Next, it describes the HMI between warfare operators and the AI-AMD system. The chapter concludes with the identification of relevant trust factors.

A. SYSTEM BOUNDARIES AND ENVIRONMENT

Two broad categories of systems are discussed in this thesis: (1) the AI-AMD overall system, which is comprised of the C2 system, sensor systems, and BLUFOR weapon systems and (2) the external systems which are comprised of the operational environment, the organizational environment, and threats. The AI-AMD C2 is expected to be operating in a System-of-systems (SoS) architecture. The integration of these multiple systems may result in emergence behaviors that are hard to predict. Figure 27 shows a context diagram of the AI-AMD C2 system and other sub-systems it interacts with that contribute to trust development between the operator and AI-AMD decision aid. The following sub-sections elaborate on the various components and interactions.

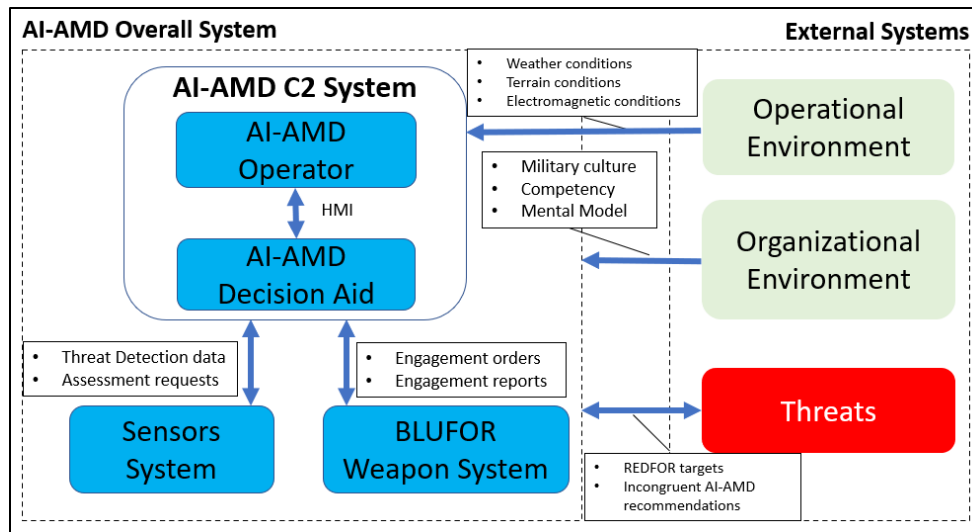


Figure 27. AI-AMD System Context Diagram.

1. AI-AMD C2 System

The C2 system consists of the AI-AMD operator working together with the AI-AMD decision aid as a teammate. The C2 system is expected to act as the central node to fuse intelligence data from the sensors system and coordinate with the BLUFOR weapon systems through engagement orders/reports into an integrated fire control network. This integrated network would provide the individual ground operators the agility necessary to support seamless operations through collaborative and shared situation awareness (SA). The C2 system is the most essential component in the entire AI-AMD overall system as it is the central authority to coordinate the entire kill-chain from collecting sensor data to engaging and re-engagement targets. Hence, the HMI between the operator and decision aid is critical. As illustrated in Chapter III on the review of the operational viewpoints and functional analysis of the AI-AMD decision aid, a total of 17 kill chain functions must be executed and coordinated with other peer systems to complete the entire kill chain. As discussed previously, a fully autonomous AI-AMD system will not be viable anytime soon. As such, the success of the C2 system is dependent on how well the operator and the AI-AMD decision aid work together and a trusting relationship is key to enable this.

a. AI-AMD Operator

As discussed in Chapter III on the review of threat scenarios, the evolving threats and technology have driven the need for a faster and more decisive response against a set of wide-ranging threats. The operator is no longer expected to deal with a single threat at a time. Instead, the operator is expected to manage a pool of resources against multiple threats simultaneously. This has driven the role of the AMD operator to change from a traditional manual controller of the system to a supervisory controller. Some of the functions that humans traditionally performed are now re-delegated to the AI-AMD. These functions include those that require psychomotor, sensory, and cognitive skills (Hawkley, Mares, and Giammanco 2005; Hopkin 1992). According to Hawkley, Mares, and Giammanco (2005), the job of a supervisory controller differs from that of a traditional manual operator.

Also, the operator is subjected to an operational environment that introduces flux and ever-changing information from electromagnetic interference, weather terrain conditions. These will create additional stressors to the operator, demanding additional cognitive workload to deal with these dynamics.

Last, as discussed in Chapter II on the review of trust factors, the organizational environment that the operators are in will influence their level of competency, and mental model shaped by the organizational culture. These factors will be identified and studied in the latter part of this chapter.

b. AI-AMD Decision Aid

The AI-AMD decision aid in this thesis is defined as an automation system that has a smart computer agent that can perform tasks with human-like intelligence that are attributed with the intent of the operator (U.S. DOD 2018). As discussed previously, one of the most promising solutions to the threat scenarios presented is the adoption of an AI-AMD system that relies extensively on automation operations. In this thesis, it is expected that the AI-AMD decision aid was built to fulfill functions that operators do poorly or at a high workload cost, as well as to augment performance in areas where operators have constraints. The AI-AMD decision aid is not designed to replace the human operator but

to augment the operator in accomplishing the mission tasks. The AI-AMD decision aid and operator are treated to work together as “teammates” with a shared goal of accomplishing the mission tasks. An important theme that arises from this notion of “teammates” is the functional allocation between human operators and AI-AMD decision aid which will be discussed later in this chapter.

2. Sensors

Sensors such as radar provide the AI-AMD C2 system with the intelligence preparations and targeting data needed to direct fires onto incoming targets. The sensors are needed to provide threat warning and situational awareness to maximize the potential capabilities of the AI-AMD system beyond the line of sight targeting and fires. Only when the AI-AMD C2 system is made known of the incoming threat, then the AI-AMD BLUFOR weapon system can be better employed for the appropriate courses of action (COA) to engage the target. Interactions between the sensors and the AI-AMD C2 system include incoming target coordinates, type and classification, and metrological data that would support the targeting of the system (Tay 2020).

3. BLUFOR Weapon System

The AI-AMD C2 System is expected to interact and coordinate with BLUFOR weapon system to engage the target. Some of these systems include the U.S. Army Patriot system, Terminal High Altitude Area Defense (THAAD) as well as non-kinetic systems to provide layered protection against aerial threats. These systems will exchange readiness reports, engagement orders, and reports, and battle damage assessment (BDA) on their targets with the AI-AMD C2 system.

4. Operational Environment

The AI-AMD system is subjected to external environmental conditions such as weather, electromagnetic traffic, and terrain. These conditions can be treated as unidirectional inputs to the AI-AMD system which affects the performance of the system. The environmental conditions will be dynamic that demands flexibility and ability by the

operator and the AI-AMD decision aid to deal with them. The operator and the AI-AMD decision aid must be trained and designed to operate in this environment for effective HMI.

5. Organizational Environment

The consideration of the organizational environment is included as trust does not develop in a vacuum and is shaped by the individual and cultural context. Individual differences such as the operator's competency level and mental model of the system are included in the individual context. The cultural backdrop influences trust through societal norms and expectations of automation and autonomous systems. These factors as previously discussed in Chapter II have an influence on trust development and are included in this study.

6. Threats

First, as discussed in Chapter III on the review of threats scenarios, the threat interaction includes a wide range of hypersonic and cruise missiles, as well as UAVs and fighter aircraft attacks, which are inputs into the AI-AMD system. These threats have become more complex and harder to destroy due to their increased operational range and velocity. This translates to a much shorter response time demanded of the operator to execute the entire kill chain to destroy the target. As such, effective HMI enabled through a trusting relationship between the operator and AI-AMD decision aid is needed to fulfill this requirement. Also, the methodology and training data used to develop the AI algorithm for the decision aid is important in building a trustworthy system. However, this will not be included as it is beyond the scope of this thesis.

The second type of threat arises in a situation when the AI-AMD decision aid recommends an incongruent recommendation from the operator's CONOPs and TTPs. This situation creates a temporal trust deficit where the operator anticipates that the AI-AMD decision aid would propose a COA that is aligned to what the operator would do. The operator will likely distrust the AI-AMD system and not adopt the recommendation which may be necessary under that situation. This type of threat situation requires a further analysis of the HMI.

B. TRUST CONCEPTUAL FRAMEWORK AND DEFINITIONS

This section presents a conceptual framework for defining trust between human warfare operators and AI-AMD decision aids. Figure 28 illustrates this conceptual framework. The y-axis represents the level of trust human has for the AI-AMD decision aid while the x-axis represents the capability of the AI-AMD decision aid. The green dotted 45° line shows the optimum trust level or calibrated trust line, where trust corresponds to AI-AMD capabilities, resulting in proper utilization (Lee and See 2004). Over-trust is represented by the region above the optimum trust level line, when trust exceeds system capabilities, resulting in misuse. The zone below the ideal level of trust shows distrust, where trust falls short of system capabilities, resulting in disuse. In this thesis, it is assumed that there exists some level of default trust (shown as the yellow box), the goal is to develop a strategy to improve the trust given an AI-AMD decision aid capability. Trust violations might occur that reduce the operator's trust when operating with the system. Hence, trust repair actions must be incorporated to maintain the trust at an optimum level.

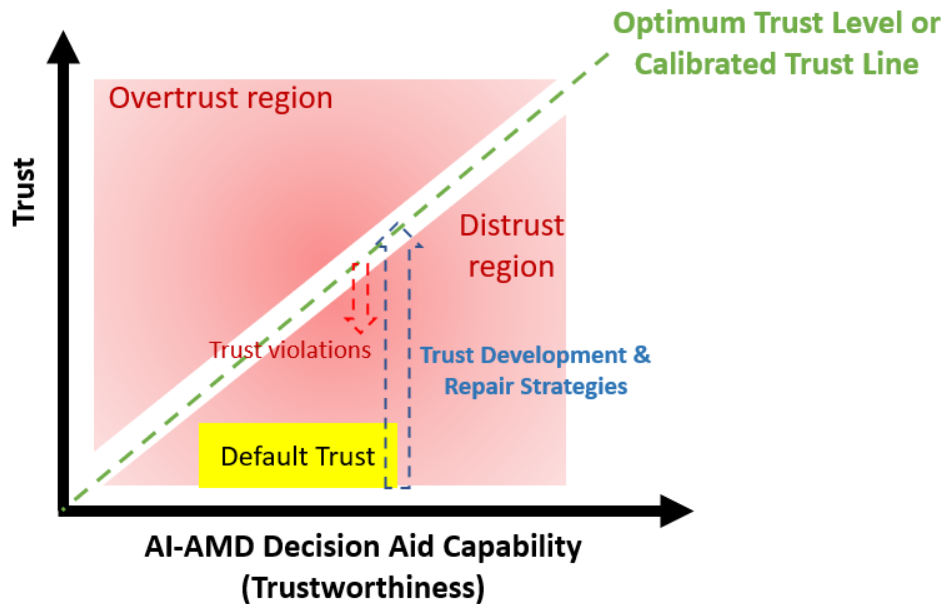


Figure 28. Trust between Human Operator and AI-AMD Decision Aid.

1. Trust

Trust refers to the operator's attitude that an AI-AMD decision aid will help to achieve the operator's goals of accomplishing the mission of destroying incoming threats promptly in an operational environment that is characterized by uncertainty and vulnerability. This definition of trust indicates that it is an emotional assessment of what the operator believes or perceives of the AI-AMD decision aid. To positively influence trust, the operator must personally see and feel the positivity of the AI-AMD decision aid actions. Hence, the inclusion of subjective factors that were previously reviewed must be included in the analysis of trust factors which will subsequently inform the strategy to improve trust.

As noted earlier, the AI-AMD operations involve a lot of uncertainties and vulnerabilities arising from the environmental factors of weather, electromagnetic interferences, and terrain as well as the nature of the ever-evolving threats. The operator will anticipate that the AI-AMD decision aid performs as per "contracts" to deal with these uncertainties and vulnerabilities. In this context, the contracts would be the perceived functions or tasks that the AI-AMD decision aid is supposed to carry out as well as the desirable performance of carrying out these functions or tasks.

2. AI-AMD Decision Aid Capability and Trustworthiness

In the topic of trust, it is important to distinguish between trust from being trustworthy and system trustworthiness. Trust as defined earlier is an attitude of the operator while trustworthiness is a property of the AI-AMD decision aid. The AI-AMD decision aid is trustworthy to some contracts if it can uphold the contracts that the user desires through its capabilities. The capabilities of the AI-AMD decision aid include the speed, accuracy, predictability, reliability of performing the functions or tasks as well as the ability to provide feedback to the operator. This thesis will not study how to improve these capabilities to improve trustworthiness. Instead, this thesis explores means to improve the operator's perception of these capabilities to improve trust.

3. Default Trust

As noted earlier in the review of trust definitions, the concept of default trust is useful as it defines the starting point of the trust level between a human operator and the AI-AMD decision aid. This will have an impact on the strategy needed to develop, maintain, and repair trust. Figure 28 illustrates the default trust level with a yellow box. In this thesis, the level of default trust is assumed to be in the distrust region such that the level of trust falls short of the AI-AMD decision aid capabilities leading to disuse which is undesirable.

4. Optimum Level of Trust/Calibrated Trust, Trust Violations and Repair

The optimum amount of trust or calibrated trust occurs when trust meets the system capabilities, as represented by the green dotted line in Figure 28. It is in the interests of the stakeholders to ensure that the trust in any system is kept at the optimum level for high performance. As noted earlier, that the initial trust an operator has on a new system would be below the desired level regardless of the amount of experience he has on other similar systems. Hence, there must be a deliberate strategy to calibrate the trust level to the desired level.

In addition, as noted earlier, trust is dynamic and it changes with time as it is affected by many factors interacting with each other continuously (Hoffman 2017; Chien et al. 2014). For instance, we would expect trust to increase as the operator uses the system for a long time and become reliant on it. This is undesirable as it may result in catastrophic failure as we know that no automation system is perfect, and it can fail. On the other hand, we would expect instances of trust violations with various consequences, which would reduce the operator's trust in the AI-AMD decision assistance. Errors, damages, time loss, and inefficiency from the AI-AMD decision aid are examples of trust violations. Trust is likely to fluctuate over time. Hence, this suggests that the strategy should have a way to estimate the optimum trust level and ensure that the trust is kept at that level during the operator's tenure as an AI-AMD operator.

C. HUMAN-MACHINE INTERACTIONS BETWEEN HUMAN OPERATORS AND AI-AMD DECISION AIDS

This section describes the HMI between human operators and AI-AMD decision aids to deepen the understanding of the AI-AMD C2 system. This provides insights into the factors influencing trust. This section describes the functional allocations between operators and an AI-AMD decision aid. Next, a characterization of the HMI based on the functional allocation is presented. Next, the section discusses design constructs across the functions. Finally, the section concludes with a discussion on the implications of the proposed HMI on trust.

1. Functional Allocations between Operator and AI-AMD Decision Aid

HMI is defined as the interactions and communications between an operator and AI-AMD decision aid via a user interface. As noted earlier, the interactions between humans and the AI-AMD decision aid system should be studied at a functionality level (across the multiple functions) as the human operator interacts in executing a function/task instead of the entire system (Lee and See 2004). As such, the analysis of HMI begins by allocating functions between operators and the AI-AMD decision aid.

The principles adopted in functional allocations are as follows (Save, Feuerberg, and Avia 2012):

- The AI-AMD C2 system can support multiple functions, each with a distinct LOA.
- The LOA was chosen by taking into account human performance parameters such as cognitive workload.
- The automation level was determined by ensuring that the human operator has a set of coherent tasks to perform for meaningful active participation in the entire operation. This will help to ensure that the operator has the same SA as the AI-AMD decision aid.

Hence, before allocating functions to either human operator, or AI-AMD decision, or both, the level of automation across all functions must be determined. The 10-point scale

of LOA by Parasuraman, Sheridan, and Wickens (2000) was adopted (see Table 1). For instance, for a function with a score of 10, that function is fully allocated to the AI-AMD decision aid as the computer does and decides everything, ignoring humans. For a function with a score of 1, the function is fully allocated to the operator as the operator performs all decisions and actions. For a function with a point scale between 2 to 9, which is very common, the functions are shared between the operator and AI-AMD decision aid. However, the ultimate responsibility of the decision made still lies on the operator as he makes the final decision to execute a function.

In this thesis, the level of automation of each of the kill chain functions was determined based on two factors: (1) decision risk and (2) cognitive workload. This consideration provides a more balanced approach in determining the level of automation by considering the negative impacts and human performance ability.

a. Decision Risk

Risk is defined as the product of an event's likelihood and impact (U.S. DOD 2017). Decision risk is a significant consideration when determining the LOA of a function. Ethically, there is still a consensus that high-risk activity should be undertaken by humans because of the accountability of outcomes. For example, it is proposed that a high-risk function/task be allocated a lower level of autonomy where the operator maintains the situation awareness and has full accountability of the decision and action taken.

Jones et al. (2020) performed the risk analysis based on the risk assessment presented in the DOD Risk Issues and Opportunities Guide, and the decision risk of each kill chain function was adopted (DOD 2017). The decision risk associated with each function is determined based on its likelihood and impact. The "risk value" can then be calculated by considering both likelihood and impact. The risk assessment value can then be calculated based on the criteria shown in Table 13. The kill chain functions, their descriptions, decision risk assessment, and values are given in Table 14.

Table 13. Risk Assessment Criteria. Adapted from DOD (2017).

Color	Risk Assessment	Risk Value
Green	Low	< 6
Green	Moderate-Low	< 12
Yellow	Moderate	< 20
Red	High	< 26

Table 14. Summary of Kill Chain Functions and Decision Risk Values. Adapted from Jones et al. (2020).

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Assessment	Decision Risk Value
Find	Collect Data	Gather sensor data from all the sensors in the network of systems. It also involves pre-processing and storing of data	Low	1
	Accept Initial Detection	Fuse massive amount of data to detect anomaly	Low	2
	Identify Emerging Target	Retrieve similar cases and match with collected data to confirm it as an anomaly	Low	3
Fix	Request Further Information	Trigger the sensors to continue gathering more sensor data on the emerging target	Low	1
	Classify Target	Decide on target type from data mining knowledge base	Moderate-Low	8
	Locate Target	Monitor in space and time and calculate the target coordinates	Moderate-Low	8
	Validate Detection	Predict the trajectory and intent of threat and confirm it as a threat	Moderate-Low	11
Track	Request Updated Target Track	Trigger the sensors to track the target	Low	5
	Validate Target	Provide precision coordinates meeting requirements of BLUFOR weapon systems	Moderate-Low	8
	Assess Blue Proximity	Combine location data of BLUFOR weapon systems against target locations and BLUFOR readiness data	High	21
Target	Nominate Engagement Options	Assess utility (capability) and readiness of BLUFOR weapon systems	Low	5
	Prioritize Targets	Rank targets based on threats level. Assess both probability and utility of threat knowledge	Moderate	14
	Select Attack Option(s)	Perform BLUFOR weapon system to target matching. Assesses both probability and utility of Course of Actions (COAs) by	High	22

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Assessment	Decision Risk Value
		matching the BLUFOR readiness, locations against target locations and types.		
Engage	Issue Orders	Send orders to individual BLUFOR weapon systems to lock-on and be ready to engage specific targets.	Low	5
	Send Fire Command	Send specific fire orders to individual BLUFOR weapon systems to shoot their allocated targets	Moderate	15
Assess	Assess Target Status	Monitor and project threats and BLUFOR weapon systems	Moderate-Low	7
	Authorize Re-attack	Send specific fire orders to individual BLUFOR weapon systems to re-shoot their allocated targets	Moderate	19

b. Cognitive Workload

The cognitive workload is the “amount of physical and mental resources needed to carry out a set of tasks in an environment” (Evans and Fendley 2017). A practical approach to functional/task allocation would be based on performance (Fitts 1951). It is reasonable that for functions with a high workload to be allocated to machines either partially or fully as humans are likely to not perform as well (Jordan 1963). While task overload seems intuitive that leads to function being automated, underload can also be problematic. Warm, Parasuraman, and Matthews (2008) for example, found that maintaining efficient visual attention to a display with little action is difficult and stressful even for highly driven individuals. Hence, more deliberate consideration of the function is warranted.

In this work, the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) as outlined in Hart (2006) was adapted to determine the workload of the kill chain functions. The NASA-TLX is a subjective cognitive workload assessment tool comprised of six somewhat independent variable subscales: mental, physical, temporal demands, frustration, effort, and performance (Hart 2006). The rating scales and definitions of these six variables are given in Figure 29. Each of the variables is scored from zero to 100 with a 5-point scale. Each function’s NASA-TLX score is calculated by averaging the scores for each of the six subscales (Evans and Fendley 2017).

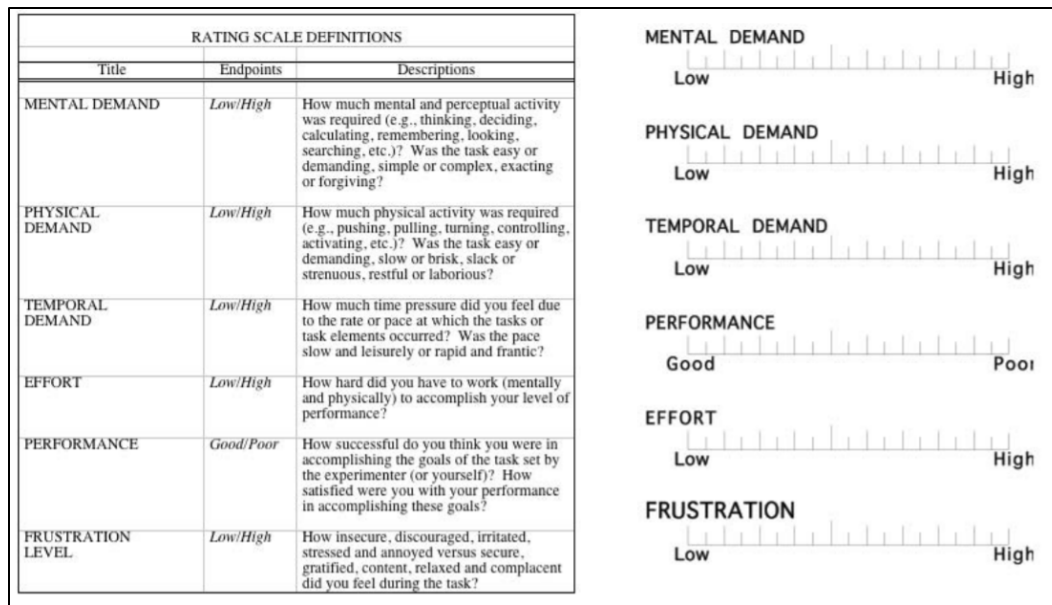


Figure 29. NASA-TLX Rating Scale and Definition. Source: Hart (2006).

An analysis was carried out across the 17 kill chain functions to determine the rating of two of the variables: mental demand and temporal demand. Physical demand was not included as all the kill chain functions are sedentary in nature that require little physical activity. All functions can be assumed to be equally undemanding. The other three were not included as they required real-life experiences with the operational system. The overall rating of the cognitive workload of each kill chain function was determined by averaging the three ratings. A summary of the ratings is given in Table 15. A similar scale as given in Table 13 was adopted to categorize the cognitive workload as shown in Table 16.

Table 15. NASA-TLX Rating across 17 Kill Chain Functions.

F2T2EA	Kill Chain Functions	NASA-TLX Variable Score		Overall Score	Cognitive Workload
		Mental Demand	Temporal Demand		
Find	Collect Data	85	50	68	Moderate
	Accept Initial Detection	90	60	75	High
	Identify Emerging Target	95	75	85	High
Fix	Request Further Information	30	60	45	Low
	Classify Target	100	75	88	High
	Locate Target	80	75	78	High
	Validate Detection	85	80	83	High
Track	Request Updated Target Track	30	60	45	Low
	Validate Target	85	85	85	High
	Assess Blue Proximity	90	85	88	High
Target	Nominate Engagement Options	60	85	78	Moderate
	Prioritize Targets	80	85	83	High
	Select Attack Option(s)	80	85	83	High
Engage	Issue Orders	50	90	70	Moderate
	Send Fire Command	40	95	68	Moderate
Assess	Assess Target Status	80	95	88	High
	Authorize Re-attack	50	100	75	High

Table 16. Cognitive Workload Criteria.

Color	Cognitive Workload	Rating
Green	Low	> 0
Green	Moderate-Low	> 49
Yellow	Moderate	> 64
Red	High	> 74

2. Proposed Level of Automation (LOA) and Functional Allocation

The LOA of the respective functions was determined using decision risk value and cognitive workload as well as considering the need to ensure that human operator has a set of coherent tasks to perform for active participation in the entire operation through a shared situation awareness with the AI-AMD decision aid. Table 17 summarizes the decision risk value, cognitive workload, proposed LOA, functional allocation, and its rationale. As part of the design to ensure that the operator maintains the same situational awareness as the AI-AMD decision aid, the operator is designed to be “in-the-loop” at every step of the F2T2EA kill chain.

The kill chain functions are allocated to two parties: AI-AMD decision aid and team, where the team indicates that both AI-AMD decision aid and operator work together in accomplishing the tasks. It is argued in every step of the kill chain functions; automation must be present to help the operator perform those tasks that are cognitively demanding. This will help to shorten the entire decision-making process to better respond to the ever-evolving threats. As such, a “team” is desired rather than the operator performing a function alone.

Table 17. Summary of Kill Chain Function Allocations.

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Value#	Cognitive Workload	Proposed LOA	Functional Allocation	Allocation Rationale
Find	Collect Data	Gather sensor data from all the sensors in the network of systems. It also involves pre-processing and storing of data	Low	Moderate	10	AI-AMD Decision Aid	Initiating the sensors to collect data can be fully autonomous with little risk to the operator. Looking through the data and understanding the data requires the operator's attention. It is better for the AI-AMD decision which can process these data faster and more consistency.
	Accept Initial Detection	Fuse massive amount of data to detect anomaly	Low	High	10	AI-AMD Decision Aid	AI-AMD decision aid is more suited to perform this function as it is demanding on cognitive workload. Also, this is a low-risk function. Hence, a LOA of 10 is preferred.
	Identify Emerging Target	Retrieve similar cases and match with collected data to confirm it as an anomaly	Low	High	7	Team	LOA of 7 was proposed to ensure that the operator maintains an oversight of the entire "Find" step. This helps to maintain the operator situation awareness.
Fix	Request Further Information	Trigger the sensors to continue gathering more sensor data on the emerging target	Low	Low	10	AI-AMD Decision Aid	This function is a C2 function carried out by the system to request more information. Also, the risk of this step is low. As such, it is desired for it to be automatically carried out.
	Classify Target	Decide on target type from data mining knowledge base	Moderate-Low	High	7	Team	This step requires a high cognitive workload as it requires the emerging threat to be check against the threat classification database. Also, there is a moderate-low risk as a wrongly classified threat could have a significant consequence. Hence, a LOA of 7 is preferred.
	Locate Target	Monitor in space and time and calculate the target coordinates	Moderate-Low	High	7	Team	Under a time-stressed scenario, it is challenging to accurately calculate and determine a target's location that is ever changing. Also, there is a moderate-low risk as a wrongly located target could have a

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Value#	Cognitive Workload	Proposed LOA	Functional Allocation	Allocation Rationale
							disastrous effect on position, navigate and time needed for response. Hence, a LOA of 7 is preferred.
	Validate Detection	Predict the trajectory and intent of threat and confirm it as a threat	Moderate-Low	High	7	Team	The operator has to decipher and fuse several information of the target initial detection, classification and locations to confirm a target is a threat. It is challenging to do it quickly and accurately. The risk involved is considered moderate-low as a wrongly validated target could have a catastrophic effect on BLUFOR. Hence, a LOA of 7 is preferred.
Track	Request Updated Target Track	Trigger the sensors to track the target	Low	Low	10	AI-AMD Decision Aid	This function is a C2 function carried out by the system to request more information as the sensors track the target after new detections are locations have been established. Also, the risk of this step is low. As such, it is desired for it to be automatically carried out.
	Validate Target	Provide precision coordinates meeting requirements of BLUFOR weapon systems	Moderate-Low	High	7	Team	This function requires the operator to fuse the target tracks and intent against the BLUFOR weapon systems. It is stressful to perform it accurately. The risk of this step is moderate-low because there are several critical follow-on functions where false alarms should be kept at a minimal.
	Assess Blue Proximity	Combine location data of BLUFOR weapon systems against target locations and BLUFOR readiness data	High	High	5	Team	This function requires merging both the threats data and BLUFOR weapon systems locations. It is challenging as there are numerous weapon systems deployed with a fast-moving target. Also, this is a high-risk function as any miscalculations could have catastrophic effects on BLUFOR. Hence, a LOA of 5.

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Value#	Cognitive Workload	Proposed LOA	Functional Allocation	Allocation Rationale
Target	Nominate Engagement Options	Assess utility (capability) and readiness of BLUFOR weapon systems	Low	Medium	10	AI-AMD Decision Aid	This function requires the operator to list out the viable options by merging the data on the availability and locations of BLUFOR weapon systems against threat trajectory. This is a low-risk function because it just involves listing the possibilities of BLUFOR weapon systems available to engage the threat. Also, it is assumed that the operator wants this to be carried out quickly for follow-on procedures. Hence, a LOA of 10 is preferred.
	Prioritize Targets	Rank targets based on threats level. Assesses both probability and utility of threat knowledge	Moderate	High	5	Team	This function is a critical and time sensitive step to determine the priority of targets based on their threat level presented to the operator. The risk is considered moderate as a wrong prioritization may result in the target not being engaged in time. Hence, a LOA of 5 is preferred where the operator approves, and acts based on what the AI-AMD decision aid recommends.
	Select Attack Option(s)	Perform BLUFOR weapon system to target matching. Assesses both probability and utility of Course of Actions (COAs) by matching the BLUFOR readiness, locations against target locations and types.	High	High	3	Team	This function requires the operator to quickly conduct BLUFOR weapon systems-targets matching by fusing many data. This is a high-risk function as a wrong match could render mission failure and disastrous effects on BLUFOR. Hence, a LOA of 3 is desired where the operator makes decisions based on the recommendations by the AI-AMD decision aid.
Engage	Issue Orders	Send orders to individual BLUFOR weapon systems to lock-on and	Low	Moderate	10	AI-AMD Decision Aid	This function requires the operator to prepare the orders and send to the designated BLUFOR weapon systems. This is also a low-risk function as the operator

F2T2EA	Kill Chain Functions	Kill Chain Function Description	Decision Risk Value#	Cognitive Workload	Proposed LOA	Functional Allocation	Allocation Rationale
		ready to engage specific targets.					has already selected the attack options. Hence, a LOA of 10 is preferred to automate this process once an attack option is selected.
	Send Fire Command	Send specific fire orders to individual BLUFOR weapon systems to shoot their allocated targets	Moderate	Moderate	5	Team	This function serves as a safety mechanism to launch an attack on the threat. However, this must be carried out quickly. This function has a higher risk than issue orders as this step is the last step before a weapon is launched against the threat. This has direct impact on mission success. As such, a LOA of 5 is preferred for the operator to decide on a check and balance.
Assess	Assess Target Status	Monitor and project threats and BLUFOR weapon systems	Moderate-Low	High	5	Team	This function requires the operator to constantly keep track of the target status and match against BLUFOR weapon systems locality and availability. This is a high-risk function as failure to assess this correctly will result in catastrophic effects given that the situation at this stage is even more time sensitive. Hence, a LOA of 5 is preferred where the AI-AMD decision aid helps to determine the COA for the operator to approve it.
	Authorize Re-attack	Send specific fire orders to individual BLUFOR weapon systems to re-shoot their allocated targets	Moderate	High	5	Team	This requires the operator to select an attack option quickly as the threat is impending. This function is a moderate risk as the operator has to choose a pre-determined attack option against the threat. Hence, a LOA of 5 is appropriate for this function.

Decision Risk Values are adapted from Jones et al. (2020).

3. Operator and AI-AMD Decision Aid HMI

Figure 30 shows the interactions between the operator and AI-AMD decision aid of those functions under “team” kill chain functions. The right half of Figure 30 shows the AI-AMD decision aid sub-system. The displays of the AI-AMD decision aid sub-system represent the system status (e.g., detection data, BLUFOR locations) and external environment (threats, environmental conditions, target tracks) in a manner that the operator can understand. Controls enable the operator to make direct changes to the system status and influence the external environment to achieve his objectives. The left half of Figure 30 shows the operator sub-system. Information and data ranging from threat details, BLUFOR readiness, and COAs from the AI-AMD decision aid display are first perceived by the operator. The operator will then process the information to make the decision. The operator will then make motor responses (e.g., accept/reject COA) to alter control settings to affect a system response. The vertical line between the AI-AMD decision Aid and operator represents the Person-Machine Interface (e.g., touch screen display, keyboard, computer mouse). This interface allows for the exchanges of information in both directions from operator to AI-AMD decision and vice versa.

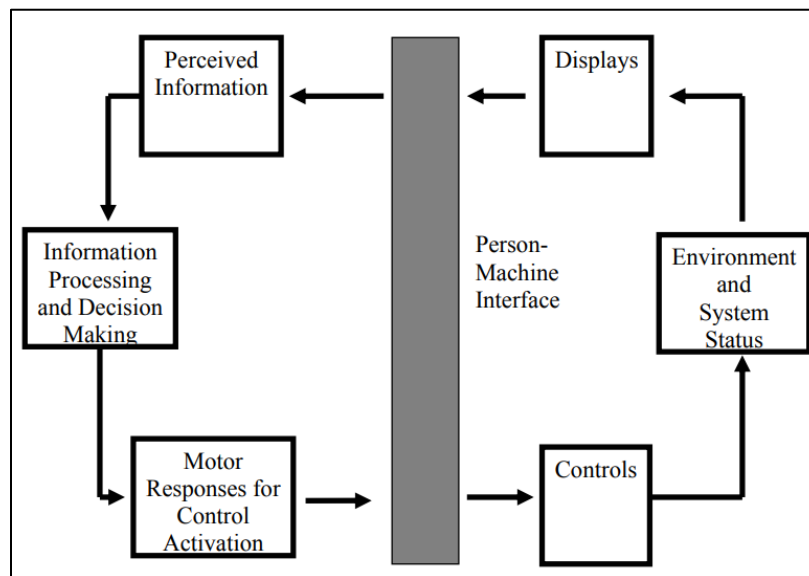


Figure 30. Schematic of the HMI of “Team” Functions. Source: Hawkley, Mares, and Giammanco (2005).

For the case of an AI-AMD decision aid allocated kill chain function, the interactions between the operator and AI-AMD decision aid are like the “team” allocated functions except that operator does not have to carry out any motor responses to effect any changes to the system as the AI-AMD decision aid will carry out the actions based on its calculations. Nonetheless, the operator will interact with the AI-AMD decision aid by observing the outputs of the AI-AMD decision aid through the display maintaining some form of situation awareness and be ready for follow-on activities.

One important observation from the study of HMI between the operator and AI-AMD decision aid is that the operator’s role has changed from a manual controller of the system to a supervisory controller. The supervisory controller is the person in charge of continuously receiving and programming information from a computer that is linked to the AI-AMD C2 process via artificial effectors and sensors (Sheridan 1992). The operator exerts control via the AI-AMD decision aid, an intermediary platform that controls the AMD process and gives information to the human (Hawkley, Mares, and Giammanco 2005).

4. Design Constructs of Operator and AI-AMD Decision Aid HMI

The operator and AI-AMD decision aid HMI is further described using the five design constructs framework as proposed by Pritchett and Feary (2011). This resolution of the HMI’s description would provide a clearer understanding of the operator’s interaction with the AI-AMD decision aid. This would then help to find factors that influence trust between the operator and AI-AMD decision aid.

a. Authority

The operator is responsible and accountable for the success and failure of the AMD operations. The operator has the full authority in making critical decisions in those kill chain functions that have been allocated as “team.” The AI-AMD decision aid operates as an “aid” to assist the operator in providing accurate and quick recommendations and COAs to hasten the operator’s decision-making process. As such, the operator would anticipate

that the AI-AMD decision aid provides recommendations or COAs that will help the operator complete the mission.

b. Team

The operator and AI-AMD decision aid can be regarded as teammates where the operator is the “leader” who is responsible and accountable for the outcome and the decision aid is the member who is supporting the overall operation. A shared mental model or situation awareness is a prerequisite for working well as a team. The operator and AI-AMD decision aid each possesses his requisite knowledge knows what information the other party possesses, and knows which information is needed by the other should he seeks it. With this shared mental model, there are clear expectations of the allocated functions and each other. A shared mental model between the operator and decision aid contributes to effective team communication and coordination. Thus, a successful human-machine team has its members possess a shared mental model for effective information sharing to help each other perform their roles better.

c. Work

The work of both the operator and AI-AMD decision aid has been defined through the functional allocation as given in Table 17. A clear understanding of the delineation of work or functions between the operator and decision aid is necessary to clarify each role and responsibility in the team.

d. Representation

The interactions of the operator and AI-AMD decision aid are dependent on the shared representations of the processes, recommendations, COAs, and decisions between the operator and decision aid. External representations, such as interfaces and displays, according to Zhang and Norman (1994), can organize cognitive behavior, affect the human’s task, and provide information that is easily comprehended and used. For the AI-AMD decision aid to be perceived as a “valuable teammate,” the representations of AI-AMD decision aid processes and recommendations must seek to improve human perceptual and cognitive processes of the AMD operations for decisive decision-making.

e. Interface Mechanism

The interface mechanism between the operator and AI-AMD decision aid facilitates information exchange and actions required by each other to carry out desired behaviors. Current interface technology such as computer displays, vocal communication, gesture communication, buttons, keyboards, and other input tools to input instructions all have limits and present HMI difficulties. A deeper study of interface design is necessary for effective HMI for trust to develop.

5. Implications of the HMI on Trust

The study on the HMI between the operator and AI-AMD decision aid within the AI-AMD C2 system suggests the following:

First, the discussion on the functional allocations based on decision risks and cognitive workload suggests that the operator and AI-AMD decision aid should be designed to operate as a “team” across most of the kill-chain functions. Hence, the factors relating to both the humans and AI-AMD decision aid must be considered in the study of the trust factor. In the study of human-related factors, the subjective factors on human’s perception of a trustworthy decision aid should be carried out. On the study of the decision aid, factors that contribute to a trustworthy system should also be carried out.

Second, the notion of a changing operator’s role from a manual controller to a supervisory controller suggests that a trustworthy decision aid is important and has a direct influence on the operator’s trust in using the system. The operator would expect that the system performs as expected to help the operator perform his role better. Also, to further help ease the cognitive workload of the operator, the external representation of the information, recommendations, and COAs from the decision aid must be easily readable and understood.

Third, the interface mechanism is a critical feature of the AI-AMD C2 system as it directly affects the interactions between the operator and decision aid. A poorly designed interface will result in a negative experience for the operator which leads to trust decrement, even if the decision aid is trustworthy.

D. TRUST FACTORS OF HUMANS AND AI-AMD DECISION AID

This section examines the five major categories of trust considerations that determine the amount of trust between the operator and the proposed AI-AMD decision aid. Figure 31 shows a summary of the trust factors considered in this thesis. First, factors associated with the operational environment which describe the context of the AMD operation system are first studied. Second, factors associated with the organizational environment where the operator and AI-AMD system have been trained and developed are studied. Third, factors associated with building a trustworthy AI-AMD decision aid are studied as a trustworthy system is a precursor for warranted trust to develop. Fourth, factors that affect an operator's performance and his attitude on the AI-AMD decision aid are studied. Last, the study of HMI suggests a "team" construct between the operator and AI-AMD decision aid. Hence, "team" factors that influence the relationship between the operator and AI-AMD decision aid are proposed.

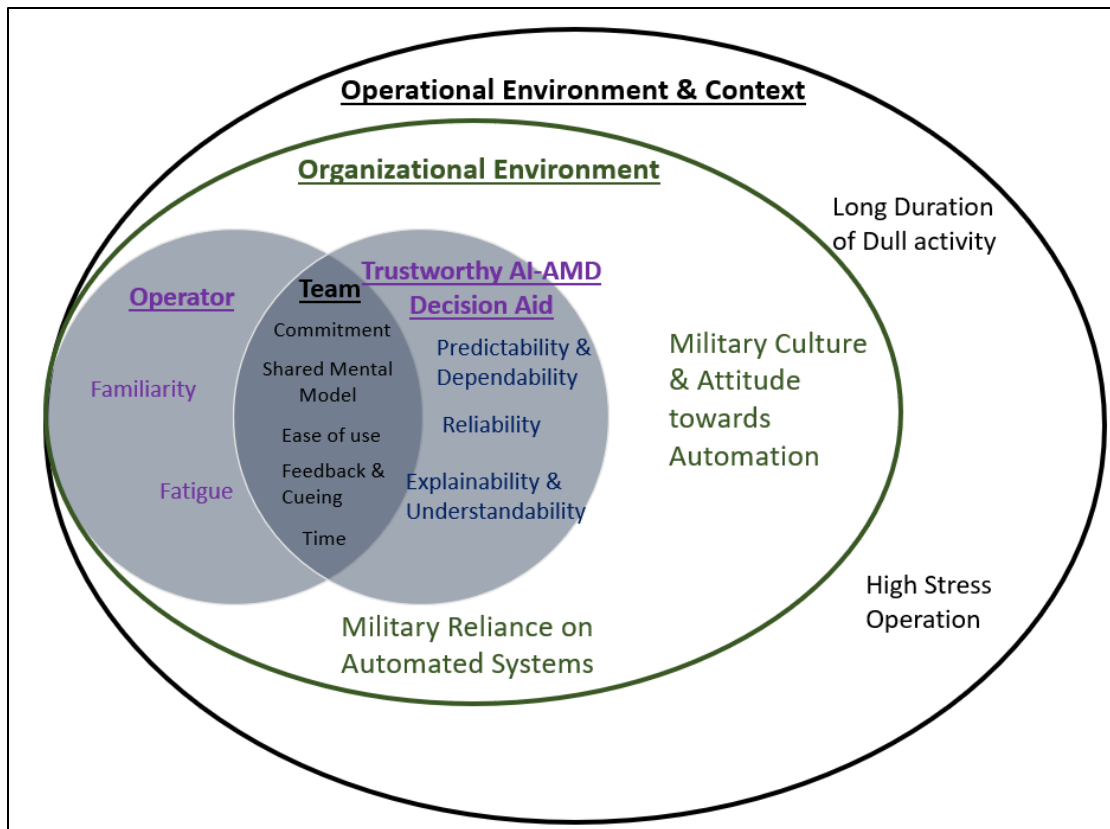


Figure 31. Summary of Trust Factors

1. Operational Environment and Context

a. Long Duration of Dull Activity

The AMD operation is defensive in nature where the operators are waiting for the sensors to detect possible threats for follow-up actions. This means that the operator is not actively engaged in the operation most of the time. This is especially true for peacetime operations where the level of threat is minimal. However, the most problematic requirement is the operation still requires the operator to maintain constant vigilance and to develop and maintain situation awareness until an anomaly happens. Additionally, there may be instances of false detections due to commercial aircraft flying past the detection zone. This may cause a decrement in trust over a prolonged period and the operator may treat a real threat as a false detection. Another undesirable situation of over-reliance may occur too. After a prolonged period of no activity, the operator may become complacent

and accepts whatever recommendations or COAs as he loses all his situation awareness of the operation.

Hence, the operator needs to remain actively engaged throughout the operation such that the level of trust is maintained. A deeper analysis of how to maintain SA throughout the operation especially the period of dull activity is needed.

b. High Stress Operation

As reviewed in Chapter III, the evolving threats cause the current AMD operations to be more time-sensitive and cognitively demanding. Moreover, the operational environment presents uncertainties and variations to the operators due to the effects of multipath, clutter, and electromagnetic jamming (Boord and Hoffman 2016). Despite these operational constraints, the operators are expected to perform well in their role as supervisory control and execute the kill chain functions without errors. These put the operators under a lot of stress (Hancock 1989). When the operators are under a lot of stress, their ability to control the system and situation decreases due to their high perception of risks (Guinot, Chiva, and Roca-Puig 2014). It is documented that stress will degrade humans' basic cognitive capacities. This will reduce the amount of information humans can process to understand the situation (Raley et al. 2004). This will reduce performance capability (Hancock 1989) and trust.

On the other hand, the operator might be too reliant on the system such that in time of the impending attack, the operator may reluctantly trust the system and accepting all the recommendations and COAs without understanding the situation and the logic. This may result in catastrophic consequences, like the U.S. Army Patriot missile incidents.

The operators' situation awareness must be maintained at the same levels as the AI-AMD decision aid to ensure that the operator is kept abreast of things to feel confident. This allows the operators to understand the situation and understand the recommendations made by the AI-AMD decision aid better, which will engender trust. A review of the HMI between the operator and AI-AMD decision aid should be carried out. The aim of this is to ensure that a sufficient LOA is designed to assist the operator to make sense of the situation

for quick decision making while also ensuring that the operator is not overly reliant on the system in the long run. In addition, methods on minimizing the number of unnecessary stressors such as false alarms, and errors should be analyzed to reduce the consequences arising from the high-stress environment.

2. Organizational Environment

a. Military Culture and Attitude Towards Automation

In general, the military culture is characterized as hierarchical, regimental, and authoritarian with strong social orders that follow rules and regulations. The military also possesses a high level of discipline (Pak et al. 2017). These attributes of a military culture developed in high-risk environments can have a substantial impact on how trust develops and maintains (Field 2020). Military servicemen, compared to civilians lack the freedom of choice. For most militaries, due to the authoritative construct of the military, many servicemen do not have the choice to choose the technological systems they operate (Field 2020). Servicemen may not have default trust in the AI-AMD decision aid if they have not been exposed to a similar system or are unwilling to take on the role of an AMD operator.

Another important aspect of the military is the strong identity forged during service. The military identity that is formed with strong attachment bonds within the “military family” may have a negative impact on trust towards AI-AMD decision aid. The strong attachment bonds are usually established under the narratives of “the perception of being involved in something significant and meaningful” (Grimell 2017). The passionate and close relationship accomplished through shared mission and hardship may develop the propensity to trust each other. However, this established relationship may impact the perception of an AI-AMD decision aid teammate as a threat to team camaraderie. AI-AMD decision aid may be viewed as a tool rather than a “teammate.” Additionally, servicemen develop a strong sense of pride and self-worth being part of this military identity. The introduction of the AI-AMD decision aid system to replace some of the operator’s tasks may result in distress and rejection in the system (Field 2020).

Hence, an analysis of how to make the AI-AMD decision aid a more “social” actor that interacts with the operators meaningfully through shared experiences should be carried

out. Also, the study of the attributes of an operator and decision aid “team” as proposed in the study of HMI, are also critical in inducting the AI-AMD decision aid into the military “family.”

b. Military Reliance on Automated Systems

To deal with the increasingly difficult operating theater, the military will continue to employ highly sophisticated and electronically advanced technologies. Specifically, automation technology has and will become increasingly important to the military forces. An automated system will likely be implemented in every corner of the military, spanning from operations, training, force protection, medical, human resource, and logistics. For instance, the U.S. military launched several initiatives to leverage AI to enhance military decision-making and operations across multiple key operational areas (U.S. DOD 2018). As such, it is expected to have more servicemen be exposed to and used automated systems in their careers. This helps to increase the automation literacy within the military. This will likely raise the level of default trust when an operator is being introduced to an AI-AMD decision aid. It is also easier to develop trust with the system as the operator already has a pre-conceived perception of the performance of a similar system.

However, there is a possibility of over-trust on the AI-AMD decision aid as more automated systems are being fielded which makes the military more reliant on automation technology. This is an undesirable consequence as the servicemen may lose their operational competencies in the long term. As such, there must be mechanisms in place to prevent servicemen from being too reliant on the decision aid.

3. AI-AMD C2 System: Operator-Related Trust Factors

a. Familiarity

Operators’ previous experience with other automated systems may contribute to a higher level of trust in automated systems in general. Operators’ experience can be gained through this career of operating different platforms as they progress in the military. The new operators to AI-AMD decision aid may start operating the system with a high level of default trust gained from operating other reliable automated systems. These operators may

welcome AI-AMD decision aid and be more willing to off-load tasks to it. For instance, the operators may have prior knowledge of the detection and identification capabilities from other systems and assume the same as the AI-AMD decision aid. However, it is crucial to highlight that even in highly reliable automation systems, operators with more expertise in an automated system may diminish a human's capacity to watch for unforeseen states. (Bailey et al. 2006). This may arise because of the operators' pre-conceived mental model of the system outputs. When these instances occur, the trust may be affected.

On the contrary, Hawley (2017) argued that the level of familiarity and interactions with automated systems is reduced as some functional tasks are allocated to the automated system. Evidence shows that when an anomaly happens, the operator will be slower to react and take necessary actions (Congress 1984). The operator loses SA when operating the system and the operator must scramble "back into the loop" to regain situation awareness and make appropriate actions. In the long run, operators may lose fundamental skillsets and proficiency in monitoring and controlling the decision aid as they get less hands-on experience.

Despite the fact that the military is fielding more automated systems and operators are expected to be more conversant with automation technology, the requirement for specific training on AI-AMD decision aid cannot be overlooked. The operational environment and system design may cause significant differences in AI-AMD decision aid as compared to other systems.

b. Fatigue due to Prolonged Monitor Supervisory Work

Operators of AI-AMD decision aids will experience active fatigue during their duties. As discussed previously, the role of AI-AMD operators will change from traditional manual control to supervisory control. An operator needs to monitor the AMD process more than ever. With automation, Operators now must monitor more indicators from the system due to an increased number of components. Prolonged vigilance on the system involves tough mental work and can be stressful. Research shows that successful supervisory control necessitates the experienced operator's ongoing and meaningful involvement with the system. (Hawley, Mares, and Giammanco 2005). As such, the

operators will experience active fatigue to maintain effective supervisory control for situation awareness even during a period of normalcy. However, the level of fatigue will increase exponentially when an anomaly is detected as the operator's cognitive workload increase because they must execute many tasks in a short time frame. This prolonged period of fatigue from a period of peace to impending attack may lead to over-dependence on decision aid as the operators may not be able to handle prolonged fatigue. This is undesirable as the operators may not be actively "monitoring" the system and approve whatever the decision aid recommends.

Hence, a deeper analysis is required to examine the methods to help the operators cope with fatigue arising from prolonged monitoring of the decision aid.

4. AI-AMD C2 System: AI-AMD Decision Aid Trustworthiness

a. Predictability and Dependability

The AI-AMD decision aid's behaviors should be predictable and dependable to positively influence trust development sustenance. It has been demonstrated that predictability serves as the foundation for trust early in a relationship (Lee and See 2004). A predictable decision aid refers to a system that matches its operator's expectations. A dependable system refers to a system that is consistent and effective in performing the allocated functions. The system must first be reliable by design to ensure predictable and consistent results. Operators' trust in the decision aid will increase if the experience with the system provides predictable outcomes based on the clearly stated HMI descriptions (Cahour and Forzy 2009; Muir 1994). However, there will be a drastic drop in trust which will result in disuse or disregard of the decision aid when the operator experiences unanticipated reactions from the system (Schaefer et al. 2016).

Additionally, the predictability and dependability of the system develop with time. As the operator spends more time with the AI-AMD decision aid, his understanding of the system processes and representations of the information will improve. These interactions over time will develop the shared mental model between the operator and the decision aid which is critical for trust development.

b. Reliability

Reliability remains one of the most studied attributes of trust in automation. Many studies have suggested that reliability has a major influence on trust development (Schaefer et al. 2016; de Vries, Midden, and Bouwhuis 2003; Lee and See 2004). Reliability refers to the degree to which the AI-AMD decision aid accomplishes the functional duties that it was supposed to do consistently throughout time. Due to the errors present, an untrustworthy decision aid will impair the operator's capacity to make decisions. This negatively affects the operator's work performance as the operator must spend more time correcting the errors or performing the tasks himself. Because the operator expects the automated aid to be accurate and reliable. Donmez et al. (2006) found that humans are likely to remember situations where the automated aid made errors than offered the right recommendations. This tendency will influence the operator's view of the automated aid reliability. This is an important observation as errors will bound to occur in automation systems as no system is perfect. As a result, it may be difficult for the operator to maintain a "proper" view of the automated aid's reliability. This may reduce trust in the system, which is unnecessary (Donmez et al. 2006).

Additionally, the human perception of the system competence such as reliability attribute plays an equally important role in trust development. According to research, operators' subjective judgments of trust in an automated system are primarily based on the operators' perceived competence of the system (Muir and Moray 1996). The authors also claimed that any indication of ineptitude in the automated system, even if it has little influence on overall system performance, will have a substantial impact on the building of trust. As a result, the operators must be kept informed of their automated aid's performance in order to maintain their views on the automated aid's reliability.

c. Explainability and Understandability

Other than the reliability of the AI-AMD decision aid, the ability to explain its decision-making process and presenting it in an understandable manner are equally important. In a Defense Advanced Research Projects Agency (DARPA)'s eXplainable AI (XAI) research program, it was found that human operators prefer explanations, even if

they degrade performance (Gunning and Aha 2019). Furthermore, the more accurately the operator understands how the decision aid works, the more trust the operator will have in it. As a result, this section delves into the ideas of explainability and understandability. Explainability refers to the details and arguments provided by the AI-AMD decision aid to make its operation clear or simple to the operator. Understandability assesses the degree to which a human operator can comprehend an AI model's choice (Arrieta et al. 2020). These definitions suggest that the understandability of the decision aid is a two-sided matter: system model explaining its functions clearly and human understanding the explanations. This section discusses two critical areas to improve explainability and understandability.

First, the design of the XAI model that determines the types, depth, and volume of explanation will be studied as the operator will prefer calibrated explanations to help in their role rather than explanations that do not add value to the operator (Gunning and Aha 2019).

Second, the interactions between the operator and decision will be studied as most conflicts and failures in complex systems are caused by these interactions issues (McDermott et al. 2018). In particular, the information representation and interface design of the decision aid will be discussed in Chapter V. These system constructs directly affect what humans perceive and thus affect trust development (Gunning and Aha 2019).

5. “Team” Related Factors

a. Commitment

Satisfaction with the system leads to commitment to the AI-AMD decision aid. In many applications, human operators' perceptions of automation are influenced by their level of pleasure. For instance, satisfaction has been demonstrated to improve trust development with combat identification (Wang, Jamieson, and Hollands 2011) and vehicle automation (Donmez et al. 2006). Arguably, the satisfaction derived from working together with the decision aid will affect the trust and relationship between the operator and decision aid. For the operator to develop trust and continue to use the system, the operator needs to perceive that the AI-AMD decision aid is a “good” and dependable teammate where it helps in the operator's work. The operators must believe that the system is specifically

developed to solve their problem (Denning 2016). Attributes such as shared mental model, transparency, and trustworthy AI-MAD decision aid are important prerequisites for operator's satisfaction as these factors affect the perception of the decision aid. Hence, there should be mechanisms to determine the satisfaction level of the interactions between the operator and AI-AMD decision aid and to identify negative aspects of the decision aid that causes dissatisfaction.

b. Shared Mental Model

The shared mental model between the operator and AI-AMD decision aid is established by the authority, intentions, assumptions, clear understanding, and limitations of each teammate's roles and responsibilities. This shared mental model needs to be constantly and actively updated and maintained so that the operator and decision aid can maintain shared SA in the area of operations and engage in coherent activities to accomplish the mission (McDermott et al. 2018). These collaborative actions will develop trust between the operator and AI-AMD decision aid over time. Hence, the design and description of HMI are critical for the operator to appreciate its limitations and how the decision aid assists the operator to perform his task better. Additionally, the explainability and understandability in the decision aid are as important to maintain the shared mental model as these contribute to how information and decisions are being conveyed to the operator.

c. Ease of Use

For effective team interactions to occur, the AI-AMD decision aid must be easy to understand and use. The scope of improving the ease of use is huge as it encompasses anthropometric, physiological, human sensory, and psychological factors. For this thesis, only the human sensory and psychological factors will be discussed. On human sensory, the vision of the operator is important as the operator spends most of his time looking at the monitors. Inappropriate design will result in operator fatigue, frustration, and inefficiency (Blanchard and Fabrycky 2011). On psychological factors, the most pertinent factor is information representation. The method, frequency, appropriateness, and timeliness of how information is presented affect how the operator's capacity to receive

and process the information. This process should not add any cognitive burden to the operator which is already operating in high-stress and cognitively demanding environment. Instead, the overall system which is AI-enabled should reduce the operator's workload. Hence, the factors on human vision sensory and information processing factors should be further explored to increase the ease of use of the AI-AMD decision aid to improve trust.

d. Feedback and Cueing

In an automated system, feedback and system cues are intimately tied to the establishment of trust (Schaefer et al. 2016). The system feedback and cues are important interactions between the operator and AI-AMD decision aid. They help to raise the operators' SA and in certain instances, help the operator to focus his attention on urgent matters. For example, cues in the form of alarms alert the operator to focus his attention on a particular anomaly, thus enhancing his job performance. As a result, the system's supporting actions are beneficial to the operator, increasing trust in the system. However, it is also important to note that the frequency and duration of feedback and cues have an impact on trust formation (Schaefer et al. 2016). Excessive feedback and false alarms cause increased alertness and may be regarded as unnecessary. In the long run, this may degrade trust. Hence, the AI-AMD decision aid should provide appropriate feedback and cues at a comfortable frequency that value-add the operator's job.

e. Time Spent Operating the System

Another important factor contributing to the performance of a team is the time spent together with the system. An operator who has used the system for a long enough period would grasp its performance, processes, and reasoning for generating recommendations and COAs. Furthermore, the operators' experience on the system allows them to recognize the system's limitations and be more tolerant of its inherent faults. This helps to develop calibrated trust in the AI-AMD decision aid. Yang et al. (2017) discovered that as the operators interact more with the automated systems that do not include a learning or adaptive element, trust in the system evolves and stabilizes. Yang et al. (2017) also found that trust increases with a reliable system and conversely decreases with a less reliable system. This observation is likely due to the interactions between the operator's initial

expectation or default trust of the system and its actual performance (Hoff and Bashir 2015a).

Hence, there is a need to increase the operator's experience with the decision aid for a trusting relationship before the operator executes a live mission. Additionally, there is a need for a feedback mechanism in the AI-AMD decision aid to be able to track interactions between the operator and decision aid to measure how trust level has changed over time.

THIS PAGE INTENTIONALLY LEFT BLANK

V. PROPOSED STRATEGY TO ACHIEVE CALIBRATED TRUST BETWEEN THE OPERATOR AND AI-AMD DECISION AID

From previous discussions on HMI and trust factors, the approach to achieve calibrated trust between the operator and AI-AMD decision aid requires not just an engineering solution on the system, it requires a holistic approach. As such, the DOTMLPF-P (Doctrine, Organization, Training, Materiel, Leadership and education, Personnel, Facilities, and Policy) solutioning framework is adopted to guide the development of the strategy in improving trust between the operator and decision aid (U.S. Joint Staff 2018). Calibrated trust occurs when the level of the operator's trust in the AI-AMD decision aid matches the AI-AMD decision aid's capability. This prevents the operators from over-trusting the system leading to over-reliance or distrust leading to disuse.

In addition, trust is dynamic and changes with time due to imperfect systems and possibly trust violations. There is a need to ensure trust remains at a calibrated level over the entire tenure of the warfighters' posting as AI-AMD operators. As such, the strategy consists of two main parts; (1) solutions to achieve calibrated trust between the operator and the AI-AMD decision, and (2) solutions to repair trust.

A. SOLUTIONS TO ACHIEVE CALIBRATED TRUST BETWEEN OPERATOR AND AI-AMD DECISION AID

Figure 32 shows the proposed strategy articulation map to achieve calibrated trust between operator and AI-AMD decision aid. The review of the trust definitions, HMI, and trust factors suggest that the strategy should focus on three key areas: (1) collective and individual humans' perception of automation technology and AI-AMD decision aid, (2) enhancing operator and AI-AMD decision aid team dynamics, and (3) trustworthiness of AI-AMD decision aid focusing on system development.

The strategy proposes three key tenets and five supporting tenets. First, military warfighters need to be informed of the true capabilities and limitations of automation technology, particularly AI-AMD decision aid. Second, the operator's training

requirements must increase to cope with a new job scope and evolving threats. Third, new requirements must be added to the AI-AMD decision aid system development to improve the system’s perceived trustworthiness. These three key tenets are supported by the other aspects of the DOTMLPF-P framework such as organization, leadership, personnel, facilities, and policy. These are discussed in the following sections.

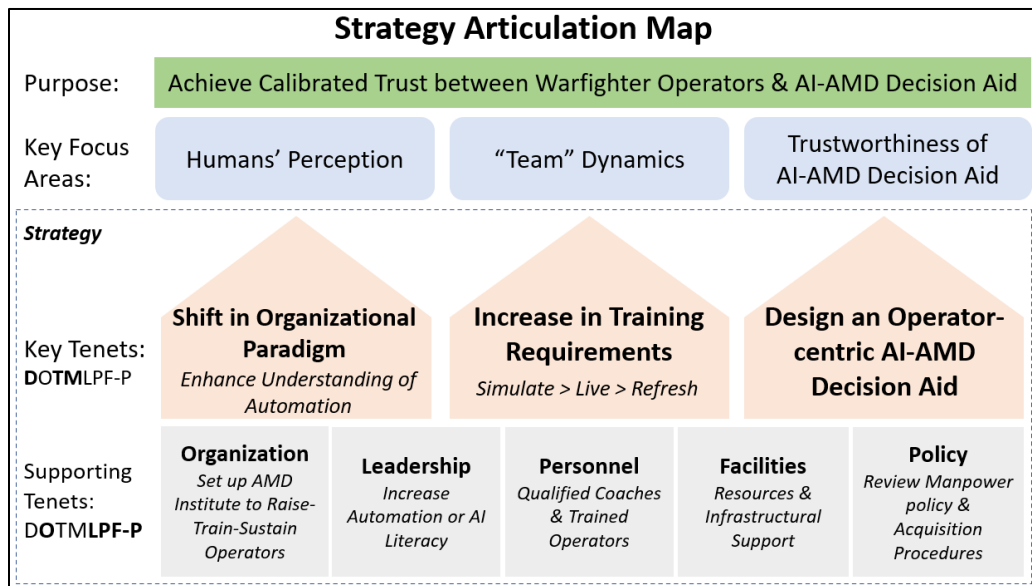


Figure 32. Strategy Articulation Map to Achieve Calibrated Trust between Operator and AI-AMD Decision Aid

1. Doctrine – Shift in Organizational Paradigm

In the military, there must be a paradigm shift in the way warfighters view automation technology and automated systems. AI has received much attention in the military as it offers tremendous opportunities to transform many domains of military operations, including warfighting, command and control, logistics, security, and maintenance, to enhance their overall operational effectiveness. As such, more automated systems are expected to be fielded. Machines or automated systems are no longer just tools; they are becoming teammates with increasing intelligence and autonomy. For operators to work well with the AI-AMD decision aid, there must first be shared beliefs on this system. The operators should then be informed of the system’s capabilities and limitations.

First, the military leaders and warfighters must be receptive to automation technology and systems to use them dutifully in operations. It was found that the introduction of an automation system to replace some of the operator's tasks may result in distress and rejection in the system (Field 2020). An organizational-level belief that automation technology and AI-AMD decision aid is necessary to achieve mission success must be created. Warfighters must not see AI-AMD decision support as a threat to existing human team cohesiveness. Furthermore, the possible ethical concerns related to AI-AMD decision aid must be clearly defined and addressed. According to a study undertaken to assess the public's default trust in an AI-enabled decision aid used for public services, confirmation that the system was not fully automatic led to higher public trust (Aoki 2021). As such, the notion that the AI-AMD decision aid being a teammate to provide quick and accurate recommendations and COAs must be emphasized to allay people's concerns that it is a fully autonomous system that kills. A great deal of work must be put into communicating and demonstrating to military leaders and tactical operators why and how the AI-AMD decision aid will benefit the AMD Command.

Second, senior military leadership must be aware that newly fielded automated or autonomous systems such as the AI-AMD decision aid rarely provide all anticipated benefits, and errors are bound to happen (Hawley 2017). This paradigm will influence all other aspects of developing and operating automated systems such as the AI-AMD decision aid. This also influences how AI is introduced into the AI-AMD system after considering its potential benefits and limitations. At the organizational level, this paradigm will create an organization's shared perspective of its military personnel on the AI-AMD decision aid capabilities and more importantly its limitations. This shared perspective on the automated systems will encourage candid conversations on the use of automation and solutions to deal with its issues. This candid communication will increase the default trust level of its personnel. At the tactical level, operators with the perception of the system's capabilities and limitations will allow themselves to calibrate their expectations of the system. Thus, the trust level of the operator on the decision aid can be calibrated appropriately without much disappointment resulting in trust decrement or over-confidence resulting in over-trust.

This shift in the organization paradigm on automation will also influence the needed support structures to enable appropriate use and trust development of AI-AMD decision aids. This will influence the training its personnel will need to leverage on the potential of the system and to overcome the inherent limitations of the system to increase the overall system effectiveness. This builds operators' confidence in the system leading to higher trust.

2. Increase in Training Requirements

“One of the hard lessons of my 35 years of experience with Patriot is that an automated system in the hands of an inadequately trained crew is a de facto fully automated system.” — Hawley (2017,4)

One of the earlier insights gained from the literature review is that automation does not necessarily make the work easier for the operator. For the AI-AMD operations where fully autonomous mode is still not a viable option in near future, the operators will continue to remain “in-the-loop” and be responsible for the outcomes of the operations (Scharre 2018). This change has also demanded a different set of skills from the operator. The nature of the work that the operator undertakes has shifted from traditional manual control to higher-level supervisory control. The operator is likely going to monitor more indications of the system status and alerts. As mentioned in Chapter IV, the nature of long periods of dull activity with short periods of high-stress situations will result in fatigue which negatively affects overall performance and operator trust in the system. Hence, this requires well-trained operators to cope with these new challenges before being deployed as AI-AMD operators.

The new training regime needs to address the following issues: disengagement, operational stress, fatigue, and system familiarity to enhance “team” dynamics. The new training regime would have to be more rigorous than before. Rasmussen, Duncan, and Leplat (1987) argued that current and future operators will have to be equipped with the same fundamental skillsets as traditional operators and those that help them in performing their roles as supervisory control. The authors cited the example of pilots who were solely taught to handle airplanes using automated flight control systems rather than being taught

to fly planes manually. These pilots would not be able to respond if their automatic control systems were to fail.

Skilled performance, according to Ericsson, Krampe, and Tesch-Römer (1993), is the result of “deliberate practice,” which they described as a protracted period of active learning in which operators continually improve his skills and ability under the guidance of a coach. Relevant feedback is provided to help the operators improve. As such, operators should be provided with experienced coaches who have operated AI-AMD decision aid systems before. Additionally, Lee and See (2004) argued that training is an important means to develop calibrated trust as this enhances operators’ understanding of the system reliability, problems, and HMI’s impacts on performance.

On “deliberate practice,” operators should undergo three phases of training: (1) simulator sessions, (2) live training with coaches to facilitate active learning, and (3) refresher training to maintain proficiency and prevent complacency as shown in Figure 33.

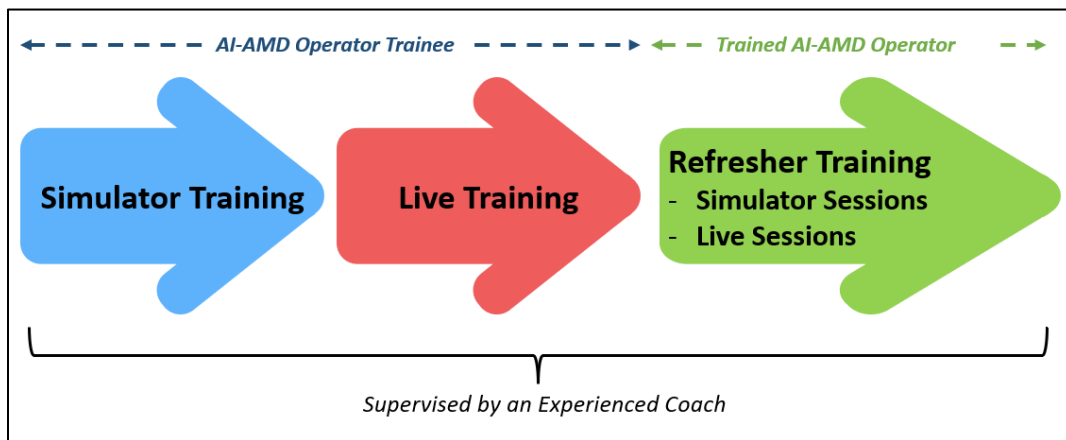


Figure 33. Proposed Training Regime for AI-AMD Operators

a. Simulator Training

Simulator sessions provide a cost-effective and low-risk training environment for operators. According to Hawley (2017), frequent simulator training is a good technique for dealing with operator skill erosion caused by out-of-the-loop familiarity. The simulator should mimic the actual operational environment and system interactions to provide good

training outcomes for the operators. The training should not just focus on “rote-drills” to get the operators qualified to undertake the unit’s operational duties (Hawkey, Mares, and Giammanco 2005). The training should instead focus on skills development, critical thinking, and problem solving under prolonged and stressed scenarios. Such training opportunities that are conducted regularly over a prolonged period allow the operators to develop a shared mental model with the decision aid. The operator will understand the capability of the system better, its limitations, and more importantly the functional allocations and each other’s responsibility. In addition, Wiener and Curry (1980) argued for the need to conduct training in manual operations followed by the intended levels of automation. They reasoned that for operators to appreciate the benefits and potential of the automated system, the operators must have some basis for comparisons.

However, simulator training alone is insufficient to retain supervisory control competency. Training for extreme situations that test the operators’ tolerance to work under fatigue and stress is difficult to replicate in simulators. For instance, system errors, faults, and random behaviors cannot be simulated. Rasmussen (1986) argued that the skills needed by the operators to perform under stressed conditions do not develop during normal training or operating modes. As such, live training should be conducted to overcome the deficiencies of simulator training.

b. Live Training

The objectives of live training are to expose the operators to (1) AI-AMD decision aid capabilities, (2) mechanisms and processes of the system in deriving recommendations, (3) inherent errors, (4) mechanisms underlying the errors, and (5) actual HMI. The operators need these experiences to fully appreciate the system to develop a shared mental model with the decision aid. Importantly, the operators must be aware of the inherent system limitations or errors such as multiple tracks on a single target to calibrate their judgement on the system’s true capability. Additionally, the HMI experience gained would allow the operators to gain a better understanding of the decision-making points that require attention. With this understanding, calibrated trust can be achieved.

In systems with similar operational domains to the AMD C2 system, Ericsson and Charness (1994) discovered that competent operators do not need to spend more time as much time monitoring the problem-solving scenario as less skilled operators to execute their duties well. They also found that highly experienced operators can maintain situation awareness (SA) with less effort. They can also re-establish SA in less time and with less effort following any disruption. This ensures that the operator is kept abreast of things to feel confident. This allows the operator to understand the situation and recommendations provided by the AI-AMD decision aid better, which will engender trust.

c. Refresher Training

Refresher training must be incorporated to maintain the operators' proficiency and trust in the system. The long duration of dull activity inherent in AMD operation will result in disengagement with the system. After a prolonged period, the operator's skills will degrade unknowingly due to the lack of practice. Like human-to-human teams, humans-to-decision aid mental model, dynamism, and teamwork will suffer because of a lack of intimate interactions. Regular refresher training and assessments must be added for the operators and decision aid to work together and maintain the mental model developed. These training and assessments should include both simulator and live sessions considering the trade-offs between budget and training realism. Simulator refresher sessions that cover those aspects of the operations that are hardly executed (second half of the F2T2EA kill chain functions) help the operators maintain their proficiency. Live refresher sessions help to increase the operators' confidence in the system when carried out in the operational environments and produced reliable results and COAs. Also, the live sessions are useful from the decision aid perspective as the interactions with the operators can be captured, processed, and added into its machine learning algorithm database to train itself. With sufficient data, this may help the decision aid to provide proactive feedback to the operator to guide his actions. This makes the system easier to use as the operators undergo more refresher training.

In all, extensive and refresher training allows the operators to gain a better understanding of the system capabilities and limitations over time. While system issues or

errors are unavoidable, familiarity with the system helps to calibrate the operators' expectations of the system and to derive their internal strategies to overcome these inherent issues.

3. Materiel – AI-AMD Decision Aid System Development

This session of Chapter V discusses the non-functional requirements (“qualities”) of the AI-AMD decision aid. Trust capabilities or “trustability,” like quality must be designed into the AI-AMD decision aid rather than added after the fact (Palmer, Selwyn, and Zwillinger 2016). These requirements include system qualities, system output characteristics, and user experience (Palmer, Selwyn, and Zwillinger 2016). Functional requirements that directly evaluate system performance, such as reliability metrics, are excluded from this thesis because they necessitate a comprehensive technical discussion that is outside the scope of this thesis. The desired decision aid is assumed to already meet the functional requirements.

a. Deliberate Functional Analysis and Allocations

Chapter IV discussed a simplistic version of the functional analysis and allocation based on risks and human cognitive workload to describe the HMI. The AI-AMD decision aid design requires much more attention and detailed consideration on the human-automation integration. Human performance and usability factors must be thoroughly considered and combined with decision aid capabilities that foster cooperative joint action and mutual trust (Chiou and Lee 2021). Effective HMI is achieved by using one's strengths to compensate for the other's weaknesses across the entire kill chain functions or activities (Ezer et al. 2019). For example, it is critical to ensure that the functional allocations do not overload or underload operators and that the remaining operators' tasks and activities are coherent (Hawkley, Mares, and Giammanco 2005). Rigorous human experiments and research with a diverse group of AMD operators must be conducted at the design stage. AMD operators' involvement at the verification and validation stage of the system development is also critical to ensure that the AI-AMD decision aid meets the real operational requirements. This continuous involvement of operators helps to build a more user-friendly decision aid that the operators want to use and have trust in it.

b. Leverage Predictive Analytics to Enhance Interactions and Predict Future Scenarios

One of the most groundbreaking AI innovations being developed now is those that permit machines to figure out what individuals are endeavoring to convey. One example of such technologies is Google Auto-Complete to suggest search terms (Goldberg et al. 1992). Google's search engine builds a personalized user model by combining the user's past behaviors and user behaviors from all over the globe. This model produces remarkable results as the search engines can correctly forecast search terms even when the user has just keyed in a few letters. While this technology may seem insignificant, it reveals that this feature can build connections with specific users over time (Hairston et al. 2012).

As such, this technology can be adapted to create an individualized AI-AMD operator model that is built upon the past interactions between the operator and decision aid, and the system database. Specifically, the AI-AMD decision aid can track operators' interactions, performance and inform them. For instance, the decision aid can provide a sound alert or visual prompt to the operators when there is a mismatch in the current and past activity logs. This can include providing a "nudge" to the operators to direct the operators' attention when the operators have lost SA after a long period of no activity from the operators. Also, the decision aid can also provide the operators visual cues to guide their actions based on their past actions. This individualized model developed over time provides the operators with a "committed teammate" that is consistent and helpful.

Second, the AI-AMD decision aid should have the capability to conduct future scenario prediction to enhance the operators' SA and to facilitate quick decision-making. The system needs to perform "simulation over live data," which involves building and running a simulation on top of the present state that uses up-to-date operational data (Palmer, Selwyn, and Zwillinger 2016). For instance, the decision aid can conduct its simulations of the likelihood of kill after the operator has selected or approved the targets options from the decision aid. The results could be presented in the form of probabilistic conclusions to the operators. This information is succinct enough to ensure that the operators stay ahead of the situation and do not suffer from the extra cognitive workload.

c. Provide Succinct and Calibrated System Feedback

As previously discussed, system feedback has been shown to contribute to the building of trust (Schaefer et al. 2016). The system feedback should be clear and succinct to reduce ambiguity (Hawkley, Mares, and Giammanco 2005). Also, the type of feedback should be calibrated to the functional task instead of a one-size-fits-all type of feedback (Yanco et al. 2016). The type of feedback should help the operator to perform a task by providing the necessary information at an appropriate instance. Examples of these include:

- A simple color (red to amber to green) indication can be displayed highlighting the progress of a task like “sending fire command.” On the other hand, text message explanation can be provided in a situation where the operators need to have a better understanding of the system such as errors.
- Another important but subtle issue is to clarify “who’s responsible” on every critical function/task of the kill chain cycle (Hawkley, Mares, and Giammanco 2005). A simple visual difference or verbal cue to the operator will suffice.
- While it is recommended to fuse information from across multiple systems within the AMD network to lower operators’ workload, the information on each system should not be neglected (McDermott et al. 2018). The AMD system involves multiple systems interacting with each other. Undesirable emergence behaviors may arise which are hard to predict. This may lower operators’ trust in the decision aid. The decision aid should be able to provide information about each of the systems under the SoS upon the operators’ requests. Information such as the individual sensors’ track of the target, BLUFOR weapon systems readiness, and its engagement status can be provided.
- Explainable AI (XAI) should be adopted to provide succinct explanations of the recommendations provided by the decision aid. The designers must adopt an operator-centric approach in developing an XAI algorithm that reports explanations that are useful for the operators. These include (1) system outputs that require explanations, (2) level of details of the explanations, and (3) the

representation of the explanations. The decision aid must provide simple explanations to allow operators to comprehend the reasoning behind its recommendations or COA in a high-stress environment (Ezer et al. 2019). These help the operators to understand the situation better which improves the operators' decision-making process.

This system feedback may enhance the trust between the operator and AI-AMD decision aid as they help to improve the operators' understanding of the situation and align the operator's mental model to the system's processes and improve

However, system feedback will inevitably increase the operator's cognitive workload because the operator has to process more information (Yanco et al. 2016). The operator should not be loaded with excessive system feedback that is deemed important. For interim design, feedback can be targeted on those "teams" allocated functions or tasks as the operators will perceive these functions to be more critical and prefer to have more information for informed decision-making. A deeper analysis and human research experiment are recommended to investigate the impact of this feedback on cognitive workload and the usefulness of this feedback. This provides resolution on the overall effect on trust as the operators' interactions with this feedback are uncertain and complex.

d. Multi-modal Interfaces to Improve Situation Awareness

The usage of different interfaces can improve operators' SA by giving effective feedback and indications. In a study by Hairston et al. (2012), it was found that the operators' attention span will be restricted if all information is provided in the same mode. For instance, simply displaying all the system outputs such as alerts, recommendations, explanations on the screen will overload the operator's visual capacity resulting in a loss of SA. Some of these system outputs can be in the form of voice (speech or noise) alerts or a combination of both voice and visual displays. The information delivered to unused senses is likely to be processed efficiently (Hairston et al. 2012). However, comprehensive designs are required to optimally distribute the modes of all system outputs that keep the operators engaged and maintain SA throughout the AMD operation.

e. Leverage Metacognition Capability to Enhance Understanding

Metacognition, often known as “knowledge of cognition” (Crowder and Friess 2011) is defined as what a system understands about itself and its own cognition. Metacognition of an AI system can theoretically enable the machine to acquire self-awareness to better understand its own processes. This would allow the AI system to self-diagnose and monitor its own uncertainty (Johnson 2021). According to Johnson (2021), the metacognition mechanism allows the AI system to avert failures by spotting indicators of impending failure and informing a human operator or moving into a safe mode of operation.

This concept of metacognition can be adopted and designed into the AI-AMD decision aid to increase calibrated trust by focusing on improving operators’ perception of the AI-AMD decision aid and how the decision aid can increase the operators’ performance.

(1) Identification and Reporting of Errors

No AI system is perfect, and errors are bound to happen. As such, whenever these errors occur, there must be mechanisms to identify system errors based on their current assessment against past data. With predictive analytics technology, the decision aid could also provide recommendations to deal with these errors. These errors and recommendations must be presented to the operators for transparency.

(2) Reporting of Projected Performance with Confidence Levels

There should be mechanisms to evaluate the projected performance of each subsystem in the AMD SoS based on their capacity to support a recommended COA. This can be accomplished by comparing current system and environment data to historical data to create an estimate of the probability of success. The data can be presented with confidence levels to help operators calibrate their trust in the system as this reporting reminds the operators that AI-AMD decision aid is imperfect and the results of the operators are dependent on a multitude of systems (Antifakos et al. 2005).

(3) Provide Estimates of AI-AMD Decision Aid Trustworthiness

There should be mechanisms to estimate the trustworthiness of the AI-AMD decision aid to inform the operators of their trust in the decision aid and to prevent over-trust or distrust. It has also been reported that decision-making improves when operators are given insights into their trustworthiness (Palmer, Selwyn, and Zwillinger 2016). To not incur additional workload on the operators in providing explicit feedback on the decision aid, Floyd, Drinkwater, and Aha (2016) proposed an inverse trust estimation method to infer how much trust operators have on decision aid. The authors proposed that the system self-estimates the trend of operators' trust (increasing, decreasing, or status quo) based on observable trust factors. This could be adopted for the AI-AMD decision aid to estimate the trust trend based on its reliability, errors rates, and interactions with the operators. This can first allow the system to self-learn and adapt its behavior to further enhance trust (Floyd, Drinkwater, and Aha 2016). Second, by displaying the trustworthiness to the operator, it enhances their own assessment of the system to calibrate its behavior towards the system.

(4) Inform AI-AMD Decision Aid's Trust on the Operator in Terms of Operator's Performance Level

There should be mechanisms to track and evaluate current and past interactions with the operators to develop an internal trust model for the operators. This model allows the decision aid to detect any operator-induced error or deviation in performance such as operator inattentiveness or fatigue. Once the decision aid detects this, it will notify the operators of the areas of concern. This feedback allows the operators to develop self-awareness through the decision aid. This increase in self-awareness helps the operators to adjust themselves for better performance.

f. Cyber Defenses

The AI-AMD decision aid, like any other AI system, should be safe and secure during AI development and during its lifetime of use (Wolff 2020). AI systems can be cyber-attacked while they are being developed, coded, trained, tested, and operationally used. It is imperative to ensure that the entire systems engineering process for developing

AI systems is protected against cyber-attacks. Cyber protection needs to continue to occur during operations as well. Any disruption to the operations by the adversary will cause unexpected results generated from the system causing operators to distrust the system.

The use of AI of providing cyber security has gained attention in the literature lately. Broadly, AI that uses machine learning algorithms can detect and respond to cyber-attacks (Sagar et al. 2019). AI can predict if incoming data is potentially harmful or not. AI could also be adopted to enhance the cybersecurity of the AI system during its development and deployment. Specifically, Zhang et al. (2021) carried out a review on 54 papers published between 2016 to 2020 on the use of AI to enhance cybersecurity. The authors highlighted that AI has been proposed to be used in user access authentication, network SA, unsafe behavior monitoring, and unusual traffic identification. The authors also presented the challenges and limitations of these methods and proposed a conceptual human-in-the-loop intelligence cybersecurity model. This concept leverages the strengths of both the humans, who are more flexible to identify new changes in computer network and AI machines who are quick in processing large amount of data to provide good recognition ability. This concept could be further developed and adapted in the AI-AMD system.

4. Organization

The AMD command must re-organize itself to strengthen raise-train-sustain of the AI-AMD operators. As previously discussed, the shift of the operator's job scope from manual control to supervisory control has indicated the need for an increase in training requirements. In addition, the training of AI-AMD operators should be supervised or facilitated by an experienced operator to provide valuable feedback. Hence, an AMD institute that is responsible for manpower planning, training programs, and logistics should be set up to institutionalize the AMD operators' training and assessment. This setup also allows deep expertise on the AI-AMD system to be developed within the AMD community as the operators can continue to enhance their skills as coaches.

5. Leadership and Education

With the fielding of more automated systems, there must be a parallel initiative to increase the automation or AI literacy within the military. The military services leadership must also be educated on the potential capabilities of automation and AI and its limitations. Leaders equipped with this knowledge can communicate frankly about automation and AI to their servicemen. This creates an environment of a shared mental model on automation and AI within the military. In addition, leaders equipped with this understanding would be in a better position in stating requirements and aid decision-making during an acquisition process of an automated system.

Education on automation throughout a serviceman's career should also be introduced. This should be conducted both formally and informally. Formal education can be included in servicemen's military courses. Informal education such as workplan seminars, leaders' sharing on specific AI applications, or case studies of a system failure can be conducted too. The objectives of this sharing are to increase military servicemen's awareness of the applications of AI and automation and the successes and failures of these applications.

At the tactical unit level, military leaders must be responsible to build a culture of reliance on AI-AMD decision aid, just like how a leader builds a close-knit human team. The AI-AMD decision aid should not be seen as a tool. It should be seen as a "teammate" just like any other operator who shares the same goal of achieving the mission. Leaders will be in the best position to influence the operators under their command with the right mindset and expectations because of the long duration of time spent together.

6. Personnel

AI-AMD coaches are needed to support the new training requirements. This personnel should have operational experience with the AI-AMD systems. AI-AMD trainees must undergo a long duration of rigorous training comprised of both simulator and live training with the system to qualify as trained operators. This builds confidence and a deep understanding of the system which helps in trust development. In addition, trained

operators who have performed well can be considered to continue as AI-AMD coaches to value-add the AMD community by contributing back the wisdom and experience gained.

7. Facilities

The new AMD institute should be resourced appropriately to conduct its missions to raise, train and sustain AMD operators. To cope with the increase in simulator training requirements, simulator training facilities should be built or restructured from an existing simulator facility. The facilities should mimic the operational environment that the operators are likely to face. For live training, additional systems must be set aside for training. Appropriate training grounds and all other resources must be secured to facilitate live and realistic training.

8. Policy

Two sets of policy changes are proposed to support the key tenets of training requirements and system development.

a. Review Manpower Policy

First, the manpower policy on the selection and the duration of operators' training must be reviewed. Given the complex nature of AMD operations, Operators for AI-AMD decision aid should be an experienced operator who has prior experience in operating similar systems. Experienced operators in this field would help the operators to learn the skills required to operate the AI-AMD decision aid. Also, experienced operators will also have a better understanding of the limitations of automated systems. They will come in with an appropriate representation of the system's capabilities and limitations. Also, the criteria in qualifying an operator for operational deployment need to be reviewed. As previously mentioned, the emphasis of "rote-drill" must be replaced with both simulator and live training results that assess the operators' adaptive decision-making skills in both normal and stressed situations. This means that the assessment period on an operator will also need to be increased for accurate assessment. This is necessary to ensure that the operator has already acquired a good understanding of the system, both the capabilities and

limitations. The initial mental model of the system will be enhanced, and it helps in the trust development on the AI-AMD decision aid.

b. Review Acquisition Processes

Second, the approach to developing a system in the military needs to be reviewed for the deployment of a trustworthy automated system. Hawley (2017) highlighted that the current linear system development procedures from requirements definition, developmental and operational testing, deployment, and operational use are often irreversible. The system design flows in one way irrespective of the downstream effects on the system development. Often, the system is evaluated during formal test events immediately before mandated project milestone review stages. However, the system at this stage is nearly completed and it is difficult to make any major changes to the system. This issue is exacerbated by the program office being overly concerned with the program schedule and budget. As such, the system may proceed with many issues and uncertainties. This is particularly concerning for the development of AI-AMD decision aid as many of the design features include human factors considerations described previously. Hence, the acquisition for such a system should be iterative in nature with changes in requirements and design solutions being accommodated as the system is being developed. This is critical as a lot of interactions with the operational users are required to develop an interactive interface during system development, and system verification and validation.

B. SOLUTIONS TO REPAIR TRUST

The key strategy for trust reparation to manage trust violations focuses on addressing the explainability and understandability of these explanations. The dynamic and complex nature of the AMD operations causes the behaviors of the decision aid to vary based on the environment and threats. Typical trust repair actions in the literature such as apology notifications may not be useful when a trust violation such as errors occur (de-Visser et al. 2016). Instead of proving an apology or any other repair action when an error occurs, XAI is adopted where the explanations of the actions are derived and given to the operators. Additionally, the explanations should be manipulated and presented succinctly to the operators for easy understanding.

As illustrated previously, XAI is adopted to provide succinct explanations of the recommendations and COA that the decision aid made. All these explanations may or may not be presented to the operators. Most of these explanations can be on a “need-by” basis as not all explanations are useful or needed by the operators. However, when a mistake or an error occurs, explanations could be provided upfront to the operators to understand the decision-making logic. The use of metacognition capability may allow the decision aid to detect its mistakes or errors and report that to the operators. By combining these capabilities, the decision aid can prompt and highlight these errors with explanations and possible solutions to deal with such errors. In addition, the explanations should be designed to be easily understood by the operators. This allows the operators to quickly comprehend the situation and reasons for errors and thus quickly re-establishing back the trust in the system. Hence, when mistakes are made, a clear and succinct explanation of the decision aid’s actions improves the knowledge of the operators. This increases or maintains rather than decreases trust.

VI. CONCLUSION AND FUTURE RESEARCH

This thesis first presented an extensive study of the definitions of trust and HMI. The thesis also reviewed the AI-AMD conceptual models, threat scenarios, and kill chains. The concepts of trust and HMI were adapted and applied to AMD operations to better understand trust and HMI between the operator and AI-AMD decision aid. This thesis sheds insights into the identification of trust factors that are relevant for operators and AI-AMD decision aids. The concept of trust was studied by understanding the system context of operational and organizational environments and the interactions between operators and decision aids. This eventually led to the adoption of the U.S. military DOTMLPF-P solutioning framework to provide a holistic approach to improve calibrated trust between operators and future AI-AMD decision aids.

A. KEY INSIGHTS

This thesis yielded four significant insights that are applicable to other automated or autonomous systems. These insights add value to our understanding of trust between human operators and automated systems.

(1) The Use of Automation Gives Rise to a New Paradigm of Complexity in Military Operations

More automated and autonomous systems will continue to be fielded in the military to increase operational effectiveness. The adoption of these new systems has given rise to a new paradigm of complexity and challenges for military operations. In the past, soldiers and sailors had to contest with the environment and enemy. Now, they are required to use automated or autonomous systems to improve their combat effectiveness in a more challenging operational environment. However, the employment of these inherently complex systems has introduced a new challenge to military operators, who will be faced with determining how much trust to place in these systems to achieve what is intended. As described previously, neither too much nor too little trust is desirable as these will lead to dire consequences. Hence, the military now must learn how to appropriately manage and trust these systems to continue to use them effectively.

(2) Calibrated Trust vs. Trust

This thesis on trust highlights that it is not trust that matters, but calibrated trust. The military should be concerned with achieving calibrated trust rather than simply focused on increasing operator trust in automated systems indefinitely. A focus on simply increasing trust will result in over-trust, which will lead to misuse and over-reliance without a healthy level of operator skepticism. Calibrated trust is achieved when the operators' trust in the AI-AMD decision aid matches that of the system capabilities. Operator's trust that matches the system capabilities will lead to effective interactions which increased overall performance. To achieve this, deliberate efforts must be taken to increase the operators' perception of the system's capabilities or trustworthiness so that the operators' behaviors can be calibrated accordingly.

(3) Study Trust in Context

Trust should not be studied in isolation, just between the human and system. The external environment and systems will have to be considered to better characterize trust. This thesis has demonstrated that trust can be studied using systems engineering tools and processes such as operational needs, threats analysis, and context diagrams. The use of context diagrams helped to identify the key systems that interact with the AI-AMD C2 system. These interactions will help to develop a deeper understanding of trust between the operators and decision aid. As such, targeted solutions can be derived to enhance trust.

(4) Holistic Strategy is Needed to Increase Calibrated Trust

This thesis has also demonstrated that the approach to achieve calibrated trust between the operator and AI-AMD decision aid requires not just an engineering solution on the system, it requires a holistic approach. This is a result of the holistic understanding of trust from operational, organizational perspectives to the internal systems. The U.S. military DOTMLPF-P solution framework has proven to be a useful tool to frame the solutions for sustainment trust development. One should not just look at improving the system to improve trust for system adoption but also look at the organization's perspective on the technology and the existing support structures.

B. FUTURE WORK

This thesis provided a conceptual analysis of trust and HMI in an AI-AMD C2 system. While these provided invaluable insights to guide the development of a holistic strategy to improve calibrated trust, more indicative results could be achieved if human research was carried out. As reviewed, the human dimension in the AI-AMD C2 system plays a critical role in describing trust between the operator and AI-AMD decision aid. As such, two recommendations on human research are given to strengthen the understanding of trust from the human dimension. In addition, the technical aspect of the system needs to be explored to better derive the technical requirements of the system to increase its inherent trustworthiness too.

(1) Human Research to Enhance Understanding of HMI

Human study and experimentation will be required to better understand the cognitive workload of AMD operators across the full kill chain. This can be conducted with AMD operators executing the operations in a similar simulation environment. The LOA across the kill chain functions proposed in Chapter IV could be adopted to validate if it is the optimal LOA with live experimentation. Moreover, the LOA across the kill chain functions can vary in the experiment to investigate its impact on the operator's workload and performance. This optimal allocation of LOA will influence the five HMI design constructs and thus provide insights into understanding trust between the operators and AI-AMD decision aid.

(2) Human Research to Determine System Attributes that Enhance Operators' Perception of a Trustworthy System

Another human research can be carried out to investigate the system attributes that enhance operators' perception of a trustworthy system. This can be done in a form of surveys on experienced AMD operators on what they perceive as helpful in understanding the information from the system and forming a mental model of the system. To improve their performance, AMD operators can also communicate what they wish to see on the user interface for each kill chain function. This research will solicit a list of non-functional requirements on the system to enhance its perceived trustworthiness.

(3) Technical Analysis of the AI-AMD Decision Aid

In this thesis, the AI-AMD decision aid was assumed to be trustworthy. This assumed that the system-related attributes such as reliability, explainability, understandability were well-established. A deeper analysis to identify all relevant system attributes of trust and the requirement specifications is required for the development of a trustworthy system. This study should also be conducted in conjunction with the human research mentioned above to synchronize system inherent trustworthiness and human perceived trustworthiness. The inherent trustworthiness of the system must be demonstrated in ways to allow the human operators to easily understand and develop a mental model of it. This allows the operators to develop calibrated trust in the decision aid.

LIST OF REFERENCES

- Abbass, Hussein A. 2019. "Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust." *Cognitive Computation* 11 (2): 159–171.
- Abe, Genya, and John Richardson. 2006. "The Influence of Alarm Timing on Driver Response to Collision Warning Systems Following System Failure." *Behaviour & Information Technology* 25 (5): 443–452.
- Airforce-technology. 2011. "Chengdu J-20 Multirole Stealth Fighter Aircraft." Airforce-technology. Accessed 22 Apr 2021. <https://www.airforce-technology.com/projects/chengdu-j20/>.
- . 2021. "Su-35 Flanker-E Multirole Fighter." Airforce-technology. <https://www.airforce-technology.com/projects/su-35/>.
- Antifakos, Stavros, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. "Towards Improving Trust in Context-aware Systems by Displaying System Confidence." Proceedings of the 7th international Conference on Human Computer Interaction with Mobile Devices & Services.
- Aoki, Naomi. 2021. "The Importance of the Assurance that "Humans are still in the Decision Loop" for Public Trust in Artificial Intelligence: Evidence from an Online Experiment." *Computers in Human Behavior* 114: 106572.
- Ardern-Jones, Joanne, Donald K Hughes, Philip H Rowe, David R Mottram, and Christopher F Green. 2009. "Attitudes and Opinions of Nursing and Medical Staff Regarding the Supply and Storage of Medicinal Products Before and After the Installation of a Drawer-Based Automated Stock-Control System." *International Journal of Pharmacy Practice* 17 (2): 95–99.
- Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115.
- Bailey, Nathan R, Mark W Scerbo, Frederick G Freeman, Peter J Mikulka, and Lorissa A Scott. 2006. "Comparison of a Brain-based Adaptive System and a Manual Adaptable System for Invoking Automation." *Human Factors* 48 (4): 693–709.
- Beyer, Hugh. 1998. *Holzblatt and Karen Contextual design: Defining Customer-centered Systems*. San Francisco: Elsevier.

- Billman, Dorrit, Michael Feary, Debra Schreckengost, and Lance Sherry. 2010. "Needs Analysis: The Case of Flexible Constraints and Mutable Boundaries." In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 4597–4612.
- Biros, David P, Mark Daly, and Gregg Gunsch. 2004. "The Influence of Task Load and Automation Trust on Deception Detection." *Group Decision and Negotiation* 13 (2): 173–189.
- Bitan, Yuval, and Joachim Meyer. 2007. "Self-initiated and Respondent Actions in a Simulated Control Task." *Ergonomics* 50 (5): 763–788.
- Blaha, Leslie M. 2018. "Interactive OODA Processes for Operational Joint Human-machine Intelligence." NATO IST-160 Specialist's Meeting: Big Data and Military Decision Making.
- Blanchard, Benjamin S, and Wolter J Fabrycky. 2011. *Systems Engineering and Analysis, 5th Edition*. Prentice Hall Englewood Cliffs, NJ.
- Boord, Warren J, and John B Hoffman. 2016. *Air and Missile Defense Systems Engineering*. CRC Press.
- Bradshaw, Jeffrey M, Paul Feltovich, and Matthew Johnson. 2012. "Human-agent Interaction." *Handbook of Human-Machine Interaction*: 283–302.
- Cahour, Béatrice, and Jean-François Forzy. 2009. "Does Projection Into Use Improve Trust and Exploration? An Example with a Cruise Control System." *Safety science* 47 (9): 1260–1270.
- Chen, J YC, and Michael J Barnes. 2012. "Supervisory Control of Multiple Robots: Effects of Imperfect Automation and Individual Differences." *Human Factors* 54 (2): 157–174.
- Chen, JYC, and Peter I Terrence. 2009. "Effects of Imperfect Automation and Individual Differences on Concurrent Performance of Military and Robotics Tasks in a Simulated Multitasking Environment." *Ergonomics* 52 (8): 907–920.
- Chien, Shih-Yi, Zhaleh Semnani-Azad, Michael Lewis, and Katia Sycara. 2014. "Towards the Development of an Inter-cultural Scale to Measure Trust in Automation." International conference on cross-cultural design.
- Chiou, Erin K, and John D Lee. 2021. "Trusting Automation: Designing for Responsivity and Resilience." *Human Factors*: 00187208211009995.
- Clavell, Gemma Galdon, and Philippe M Frowd. 2014. "Using Societal Impact Assessment (SIA) to Improve Technological Development in the Field of Crime Prevention." International Conference on Social Informatics.

- Cohen, Marvin S, Raja Parasuraman, and Jared T Freeman. 1998. "Trust in Decision Aids: A model and its Training Implications." in Proc. Command and Control Research and Technology Symp.
- Congress, U.S. 1984. "Office of Technology Assessment, Computerized manufacturing automation." *Employment, Education and the Workplace*: 104.
- Cosenzo, Keryl A, Raja Parasuraman, Anthony Novak, and Michael Barnes. 2006. *Implementation of Automation for Control of Robotic Systems*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING
- Crowder, James A, and Shelli MA NCC Friess. 2011. "Metacognition and Metamemory Concepts for AI Systems." Proceedings on the International Conference on Artificial Intelligence (ICAI).
- Cuevas, Haydee M, Stephen M Fiore, Barrett S Caldwell, and LAuRA StRAtER. 2007. "Augmenting Team Cognition in Human-automation Teams Performing in Complex Operational Environments." *Aviation, Space, and Environmental Medicine* 78 (5): B63–B70.
- de-Visser, Ewart J, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. "Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents." *Journal of Experimental Psychology: Applied* 22 (3): 331.
- de-Visser, Ewart J, Richard Pak, and Tyler H. Shaw. 2018. "From 'Automation' to 'Autonomy': the Importance of Trust Repair in Human–machine Interaction." *Ergonomics* 61 (10): 1409–1427. <https://doi.org/10.1080/00140139.2018.1457725>. <https://doi.org/10.1080/00140139.2018.1457725>.
- de Vries, Peter, and Cees Midden. 2008. "Effect of Indirect Information on System Trust and Control Allocation." *Behaviour & information technology* 27 (1): 17–29.
- de Vries, Peter, Cees Midden, and Don Bouwhuis. 2003. "The Effects of Errors on System Trust, Self-confidence, and the Allocation of Control in Route Planning." *International Journal of Human-Computer Studies* 58 (6): 719–735.
- Denning, Peter J. 2016. "Software Quality." *Commun. ACM* 59 (9): 23–25. <https://doi.org/10.1145/2971327>. <https://doi.org/10.1145/2971327>.
- DOD, U.S. 2017. Department of Defense Risk, Issue, and Opportunity Management Guide for Defense Acquisition Programs edited by Department of Defense. Washington, DC: Department of Defense.

- Donmez, Birsen, Linda Ng Boyle, John D Lee, and Daniel V McGehee. 2006. "Drivers' Attitudes Toward Imperfect Distraction Mitigation Strategies." *Transportation Research Part F: Traffic Psychology and Behaviour* 9 (6): 387–398.
- Duhigg, Charles. 2016. *Smarter Faster Better: The Transformative Power of Real Productivity*. Random House.
- Ericsson, K Anders, and Neil Charness. 1994. "Expert Performance: Its Structure and Acquisition." *American Psychologist* 49 (8): 725.
- Ericsson, K Anders, Ralf T Krampe, and Clemens Tesch-Römer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review* 100 (3): 363.
- Evans, Dakota C, and Mary Fendley. 2017. "A Multi-measure Approach for Connecting Cognitive Workload and Automation." *International Journal of Human-Computer Studies* 97: 182–189.
- Ezer, Neta, Sylvain Bruni, Yang Cai, Sam J Heppenstal, Christopher A Miller, and Dylan D Schmorow. 2019. "Trust Engineering for Human-AI Teams." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Ezer, Neta, Arthur D Fisk, and Wendy A Rogers. 2008. "Age-related Differences in Reliance Behavior Attributable to Costs Within a Human-decision Aid System." *Human factors* 50 (6): 853–863.
- Field, Megan. 2020. "The Facilitation of Trust in Automation: A Qualitative Study of Behaviour and Attitudes Towards Emerging Technology in Military Culture." The University of Birmingham.
- Fishbein, Martin, and Icek Ajzen. 1977. "Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research."
- Fitts, Paul M. 1951. "Human Engineering for an Effective Air-navigation and Traffic-control System."
- Floyd, Michael W, Michael Drinkwater, and David W Aha. 2016. "Learning Trustworthy Behaviors using an Inverse Trust Metric." In *Robust Intelligence and Trust in Autonomous Systems*, 33–53. Springer.
- Fogg, Brian J, and Hsiang Tseng. 1999. "The Elements of Computer Credibility." Proceedings of the SIGCHI conference on Human Factors in Computing Systems.
- Galdorisi, George. 2019. "The Navy Needs AI, It's Just Not Certain Why." USNI Proceedings.

- Gao, Ji, John D Lee, and Yi Zhang. 2006. "A Dynamic Model of Interaction Between Reliance on Automation and Cooperation in Multi-operator Multi-Automation Situations." *International Journal of Industrial Ergonomics* 36 (5): 511–526.
- Goldberg, David, David Nichols, Brian M Oki, and Douglas Terry. 1992. "Using Collaborative Filtering to Weave an Information Tapestry." *Communications of the ACM* 35 (12): 61–70.
- Gottman, John M, and Robert W Levenson. 1992. "Marital Processes Predictive of Later Dissolution: Behavior, Physiology, and Health." *Journal of Personality and Social Psychology* 63 (2): 221.
- Grimell, Jan. 2017. "A Service Member's Self in Transition: A longitudinal Case Study Analysis." *Journal of Constructivist Psychology* 30 (3): 255–269.
- Guinot, Jacob, Ricardo Chiva, and Vicente Roca-Puig. 2014. "Interpersonal Trust, Stress and Satisfaction at Work: An Empirical Study." *Personnel Review*.
- Gunning, David, and David Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." *AI Magazine* 40 (2): 44–58.
- Gupta, Nitin, Ann M Bisantz, and Tarunraj Singh. 2002. "The Effects of Adverse Condition Warning System Characteristics on Driver Performance: An Investigation of Alarm Signal Type and Threshold Level." *Behaviour & Information Technology* 21 (4): 235–248.
- Hairston, W David, Jesse Chen, Michael Barnes, Ivan Martinez, Michael LaFiandra, Mary Binseel, Angelique Scharine, Mark Ericson, Barry Vaughan, and Brent Lance. 2012. *Technological Areas to Improve Soldier Decisiveness: Insights from the Soldier-system Design Perspective*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING
- Hancock, Peter A. 1989. "A Dynamic Model of Stress and Sustained Attention." *Human Factors* 31 (5): 519–537.
- Hancock, Peter A, Deborah R Billings, and Kristen E Schaefer. 2011. "Can you Trust your Robot?" *Ergonomics in Design* 19 (3): 24–29.
- Hancock, Peter A, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. "A Meta-analysis of Factors Affecting Trust in Human-robot Interaction." *Human Factors* 53 (5): 517–527.
- Hart, Sandra G. 2006. "NASA-task Load Index (NASA-TLX); 20 years later." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

- Hawkley, John K, Anna L Mares, and Cheryl A Giammanco. 2005. *The Human Side of Automation: Lessons for Air Defense Command and Control*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING
- Hawley, John K. 2017. "Patriot wars: Automation and the Patriot Air and Missile Defense System."
- Hawley, Katherine. 2014. "Trust, Distrust and Commitment." *Noûs* 48 (1): 1–20.
- Ho, Geoffrey, Dana Wheatley, and Charles T Scialfa. 2005. "Age Differences in Trust and Reliance of a Medication Management System." *Interacting with Computers* 17 (6): 690–710.
- Hoff, K. A., and M. Bashir. 2015a. "Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust." *Human Factors* 57 (3): 407–34. <https://doi.org/10.1177/0018720814547570>. <https://www.ncbi.nlm.nih.gov/pubmed/25875432>.
- Hoff, Kevin Anthony, and Masooda Bashir. 2015b. "Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust." *Human Factors* 57 (3): 407–434.
- Hoffman, Robert R. 2017. "A Taxonomy of Emergent Trusting in the Human–machine Relationship." *Cognitive Systems Engineering: The Future for a Changing World*: 137–163.
- Hoffman, Robert R, Paul J Feltovich, Kenneth M Ford, and David D Woods. 2002. "A Rose by any Other Name... Would Probably be Given an Acronym [Cognitive Systems Engineering]." *IEEE Intelligent Systems* 17 (4): 72–80.
- Hoffman, Robert R, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. "Trust in Automation." *IEEE Intelligent Systems* 28 (1): 84–88.
- Hopkin, VD. 1992. "Human Factors Issues in Air Traffic Control." *Human Factors Society Bulletin* 35 (6): 1–4.
- Huerta, Esperanza, TerryAnn Glandon, and Yanira Petrides. 2012. "Framing, Decision-aid Systems, and Culture: Exploring Influences on Fraud Investigations." *International Journal of Accounting Information Systems* 13 (4): 316–333.
- Jacovi, Alon, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

- Johnson, Bonnie. 2021. "Metacognition for Artificial Intelligence System Safety [Manuscript Submitted for Publication]." *Proceedings of the 16th International Conference on Systems (ICONS)*.
- Jones, J. Issac, Russell Kress, Adam Rahman, and William Newmeyer. 2020. "Leveraging Artificial Intelligence (AI) for Air and Missile Defense (AMD): An Outcome Oriented Decision Aid." Systems Engineering Masters Programme, Systems Engineering Department, Naval Postgraduate School. <https://calhoun.nps.edu/handle/10945/66088>.
- Jordan, Nehemiah. 1963. "Allocation of Functions Between Man and Machines in Automated Systems." *Journal of Applied Psychology* 47 (3): 161.
- Kennedy, William G, Ciara Sibley, and Joseph T Coyne. 2015. "Self-Confidence of Autonomous Systems in a Military Environment." AAAI Fall Symposia.
- Koustanaï, Arnaud, Viola Cavallo, Patricia Delhomme, and Arnaud Mas. 2012. "Simulator Training with a Forward Collision Warning System: Effects on Driver-system Interactions and Driver Trust." *Human factors* 54 (5): 709–721.
- Lamberth, Megan 2020. AI Ethical Principles: Implementing U.S. Military's Framework. Singapore: S.Rajaratnam School of International Studies, Nanyang Technological University, Singapore.
- Lee, John D, and Katrina A See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80.
- Lee, John, and Neville Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human-machine Systems." *Ergonomics* 35 (10): 1243–1270.
- Lees, Monica N, and John D Lee. 2007. "The Influence of Distraction and Driving Context on Driver Response to Imperfect Collision Warning Systems." *Ergonomics* 50 (8): 1264–1286.
- Li, Yung-Ming, and Yung-Shao Yeh. 2010. "Increasing Trust in Mobile Commerce Through Design Aesthetics." *Computers in Human Behavior* 26 (4): 673–684.
- Lin, Patrick, George Bekey, and Keith Abney. 2008. *Autonomous Military Robotics: Risk, Ethics, and Design*. California Polytechnic State Univ San Luis Obispo.
- Ma, Ruiqi, and David B Kaber. 2007. "Effects of In-vehicle Navigation Assistance and Performance on Driver Trust and Vehicle Control." *International Journal of Industrial Ergonomics* 37 (8): 665–673.
- Madhavan, Poornima, and Douglas A Wiegmann. 2007. "Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems." *Human Factors* 49 (5): 773–785.

- Matthews, Gerald, Ryan Wohleber, Jinchao Lin, and April Rose Panganiban. 2019. "Fatigue, Automation, and Autonomy: Challenges for Operator Attention, Effort, and Trust." *Human Performance in Automated and Autonomous Systems*: 127–150.
- McCabe, V. 2014. *Coming to Our Senses: Perceiving Complexity to Avoid Catastrophes*. Oxford University Press.
- McDermott, Patricia, Cindy Dominguez, Nicholas Kasdaglis, Matthew Ryan, Isabel Trhan, and Alexander Nelson. 2018. *Human-Machine Teaming Systems Engineering Guide*. MITRE CORP BEDFORD MA BEDFORD United States.
- Mercado, Joseph E, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. "Intelligent Agent Transparency in Human-agent Teaming for Multi-UxV Management." *Human Factors* 58 (3): 401–415.
- Merritt, Stephanie M, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. "I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system." *Human Factors* 55 (3): 520–534.
- Merritt, Stephanie M, and Daniel R Ilgen. 2008. "Not All Trust is Created Equal: Dispositional and History-based Trust in Human-automation Interactions." *Human Factors* 50 (2): 194–210.
- Mittu, Ranjeev, Donald Sofge, Alan Wagner, and William Frere Lawless. 2016. *Robust Intelligence and Trust in Autonomous Systems*. Springer.
- Moray, Neville, Toshiyuki Inagaki, and Makoto Itoh. 2000. "Adaptive Automation, Trust, and Self-confidence in Fault Management of Time-critical Tasks." *Journal of Experimental Psychology: Applied* 6 (1): 44.
- Muir, Bonnie M. 1994. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." *Ergonomics* 37 (11): 1905–1922.
- Muir, Bonnie M, and Neville Moray. 1996. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." *Ergonomics* 39 (3): 429–460.
- Norman, Donald A. 1984. "Stages and Levels in Human-machine Interaction." *International Journal of Man-machine Studies* 21 (4): 365–375.
- O'Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge University Press.
- Pak, Richard, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. 2012. "Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults." *Ergonomics* 55 (9): 1059–1072.

- Pak, Richard, Ericka Rovira, Anne Collins McLaughlin, and Natalee Baldwin. 2017. "Does the Domain of Technology Impact User Trust? Investigating Trust in Automation Across Different Consumer-Oriented Domains in Young Adults, Military, and Older Adults." *Theoretical Issues in Ergonomics Science* 18 (3): 199–220.
- Pak, Richard, Ericka Rovira, Anne Collins McLaughlin, and William Leidheiser. 2017. "Evaluating Attitudes and Experience with Emerging Technology in Cadets and Civilian Undergraduates." *Military Psychology* 29 (5): 448–455.
- Palmer, Gari, Anne Selwyn, and Dan Zwillinger. 2016. "The "Trust V": Building and Measuring Trust in Autonomous Systems." In *Robust Intelligence and Trust in Autonomous Systems*, 55–77. Springer.
- Parasuraman, Raja, and Christopher A Miller. 2004. "Trust and Etiquette in High-criticality Automated Systems." *Communications of the ACM* 47 (4): 51–55.
- Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39 (2): 230–253. <https://doi.org/10.1518/001872097778543886>. <https://journals.sagepub.com/doi/abs/10.1518/001872097778543886>.
- Parasuraman, Raja, Thomas B Sheridan, and Christopher D Wickens. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30 (3): 286–297.
- Pritchett, Amy, and Michael Feary. 2011. "Designing Human–Automation Interaction." In *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach*, 267–281. Ashgate Publishing Ltd.
- Rajaonah, Bako, Françoise Anceaux, and Fabrice Vienne. 2006. "Study of Driver Trust During Cooperation with Adaptive Cruise Control." *Le travail Humain* 69 (2): 99–127.
- Rajaonah, Bako, Nicolas Tricot, Françoise Anceaux, and Patrick Millot. 2008. "The Role of Intervening Variables in Driver–ACC Cooperation." *International Journal of Human-computer Studies* 66 (3): 185–197.
- Raley, Colby, Roy Stripling, Amy Kruse, Dylan Schmorrow, and James Patrey. 2004. "Augmented Cognition Overview: Improving Information Intake Under Stress." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Rasmussen, Jens. 1986. "Information Processing and Human-machine Interaction." *An Approach to Cognitive Engineering*.
- Rasmussen, Jens, Keith Duncan, and Jacques Leplat. 1987. *New Technology and the Effect of Human Error*. Wiley.

- Reichenbach, Julianne, Linda Onnasch, and Dietrich Manzey. 2011. "Human Performance Consequences of Automated Decision Aids in States of Sleep Loss." *Human Factors* 53 (6): 717–728.
- Rouse, William B, and Nancy M Morris. 1986. "On Looking into the Black Box: Prospects and Limits in the Search for Mental Models." *Psychological Bulletin* 100 (3): 349.
- Rovira, Ericka, Kathleen McGarry, and Raja Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors* 49 (1): 76–87.
- Ryan, M. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Sci Eng Ethics* 26 (5): 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>. <https://www.ncbi.nlm.nih.gov/pubmed/32524425>.
- Sagar, BS, S Niranjana, Nithin Kashyap, and DN Sachin. 2019. "Providing Cyber Security using Artificial Intelligence – A survey." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).
- Sarter, Nadine B, David D Woods, and Charles E Billings. 1997. "Automation Surprises." *Handbook of Human Factors and Ergonomics* 2: 1926–1943.
- Satterfield, Kelly, Carryl Baldwin, Ewart de Visser, and Tyler Shaw. 2017. "The Influence of Risky Conditions in Trust in Autonomous Systems." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Sauer, Juergen, Peter Nickel, and David Wastell. 2013. "Designing Automation for Complex Work Environments Under Different Levels of Stress." *Applied Ergonomics* 44 (1): 119–127.
- Save, Luca, Beatrice Feuerberg, and E Avia. 2012. "Designing Human-automation Interaction: A new Level of Automation Taxonomy." *Proc. Human Factors of Systems and Technology* 2012.
- Sayler, Kelley M. 2021. International Discussions Concerning Lethal Autonomous Weapon Systems. edited by Congressional Research Service. Washington D.C: Congressional Research Service.
- Schaefer, Kristin. 2013. "The Perception and Measurement of Human-robot Trust."
- Schaefer, Kristin E, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. "A Meta-analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems." *Human Factors* 58 (3): 377–400.

- Schaefer, Kristin E, and David R Scribner. 2015. "Individual Differences, Trust, and Vehicle Autonomy: A Pilot Study." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Schaefer, Kristin E., Edward R. Straub, Jessie Y. C. Chen, Joe Putney, and A. W. Evans. 2017. "Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-agent Teams." *Cognitive Systems Research* 46: 26–39. <https://doi.org/10.1016/j.cogsys.2017.02.002>.
- Scharre, Paul. 2018. *Army of None: Autonomous Weapons and the Future of War*. New York/London: W. W. Norton & Company.
- Scott, Kevin D. 2017. Joint Publication 3-01 Countering Air and Missile Threats. edited by U.S. Department of Defense. Washington D.C: Department of Defense, US.
- Sharpley, Sarah, Alex Stedmon, Gemma Cox, Alistair Nicholls, Tracey Shuttleworth, and John Wilson. 2007. "Flightdeck and Air Traffic Control Collaboration Evaluation (FACE): Evaluating aviation communication in the laboratory and field." *Applied Ergonomics* 38 (4): 399–407.
- Sheridan, T. B. 1988. "Trustworthiness of Command and Control Systems." *IFAC Proceedings Volumes* 21 (5): 427–431. [https://doi.org/10.1016/s1474-6670\(17\)53945-2](https://doi.org/10.1016/s1474-6670(17)53945-2).
- . 2002. *Humans and Automation: System Design and Research Issues*. Human Factors and Ergonomics Society.
- . 2019. "Individual Differences in Attributes of Trust in Automation: Measurement and Application to System Design." *Front Psychol* 10: 1117. <https://doi.org/10.3389/fpsyg.2019.01117>. <https://www.ncbi.nlm.nih.gov/pubmed/31178783>.
- Sheridan, Thomas B. 1992. *Telerobotics, Automation, and Human Supervisory Control*. MIT press.
- Sheridan, Thomas B, and William L Verplank. 1978. *Human and Computer Control of Undersea Teleoperators*. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
- Shewhart, Walter Andrew. 1931. *Economic Control of Quality of Manufactured Product*. Macmillan And Co Ltd, London.
- Soeters, Joseph L, Donna J Winslow, and Alise Weibull. 2006. "Military Culture." In *Handbook of the Sociology of the Military*, 237–254. Springer.

- Song, Lixia, and James K Kuchar. 2003. "Dissonance Between Multiple Alerting Systems. I. Modeling and Analysis." *IEEE Transactions on Systems, man, and Cybernetics-Part A: Systems and Humans* 33 (3): 366–375.
- Spain, Randall D, and James P Bliss. 2008. "The Effect of Sonification Display Pulse Rate and Reliability on Operator Trust and Perceived Workload During a Simulated Patient Monitoring Task." *Ergonomics* 51 (9): 1320–1337.
- Stanton, Neville A, Mark S Young, and Guy H Walker. 2007. "The Psychology of Driving Automation: a Discussion with Professor Don Norman." *International Journal of Vehicle Design* 45 (3): 289–306.
- Szalma, James L, and Grant S Taylor. 2011. "Individual Differences in Response to Automation: The Five Factor Model of Personality." *Journal of Experimental Psychology: Applied* 17 (2): 71.
- Tallant, Jonathan. 2017. "Commitment in Cases of Trust and Distrust." *Thought: A Journal of Philosophy* 6 (4): 261–267.
- Tay, Derek T. 2020. "A Systems Analysis on the Effectiveness of Hyper-velocity Gun System (HVPGS) for Ground-based Air and Missile Defense ," Monterey, CA; Naval Postgraduate School.
- Taylor, Gavin, Ranjeev Mittu, Ciara Sibley, and Joseph Coyne. 2016. "Towards Modeling the Behavior of Autonomous Systems and Humans for Trusted Operations." In *Robust Intelligence and Trust in Autonomous Systems*, 11–31. Springer.
- Tenney, Yvette J, William H Rogers, and Richard W Pew. 1998. "Pilot Opinions on Cockpit Automation Issues." *The International Journal of Aviation Psychology* 8 (2): 103–120.
- Toreini, Ehsan, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. "The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
- U.S. DOD. 2018. Summary of the 2018 Department of Defense Artificial Intelligence Strategy. edited by Department of Defense. Washington, DC: Department of Defense.
- . 2020, "DOD Adopts Ethical Principles for Artificial Intelligence," <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>.

- U.S. Joint Staff. 2018. *Charter of the Joint Requirements Oversight Council (JROC) and Implementation of the Joint Capabilities Integration and Development System (JCIDS)*. (Washington, DC).
- Walker, Guy H, Neville A Stanton, and Mark S Young. 2001. "Where is Computing Driving Cars?" *International Journal of Human-Computer Interaction* 13 (2): 203–229.
- Wang, Lu, Greg A Jamieson, and Justin G Hollands. 2011. "The Effects of Design Features on Users' Trust in and Reliance on a Combat Identification System." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Warm, Joel S, Raja Parasuraman, and Gerald Matthews. 2008. "Vigilance Requires Hard Mental Work and is Stressful." *Human factors* 50 (3): 433–441.
- Wickens, Christopher D, Stephen Rice, David Keller, Shaun Hutchins, Jamie Hughes, and Krisstal Clayton. 2009. "False Alerts in Air Traffic Control Conflict Alerting System: Is There a "Cry Wolf" Effect?" *Human factors* 51 (4): 446–462.
- Wicks, Andrew C, Shawn L Berman, and Thomas M Jones. 1999. "The Structure of Optimal Trust: Moral and Strategic Implications." *Academy of Management review* 24 (1): 99–116.
- Wiegmann, Douglas, Jason S McCarley, Arthur F Kramer, and Christopher D Wickens. 2006. "Age and Automation Interact to Influence Performance of a Simulated Luggage Screening Task." *Aviation, Space, and Environmental Medicine* 77 (8): 825–831.
- Wiener, Earl L, and Renwick E Curry. 1980. "Flight-deck Automation: Promises and Problems." *Ergonomics* 23 (10): 995–1011.
- Winnefield, JA, and F Kendall. 2012. "Unmanned Systems Integrated Foadmap FY2011-2036 (Reference number 11-S-36130)." *Department of Defense*.
- Wolff, Josephine. 2020. "How to Improve Cybersecurity for Artificial Intelligence." Brookings Institution's Artificial Intelligence and Emerging Technology (AIET). Accessed 18 Aug. <https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/>.
- Wright, Stephen. 2010. "Trust and Trustworthiness." *Philosophia* 38 (3): 615–627. <https://doi.org/10.1007/s11406-009-9218-0>. <https://doi.org/10.1007/s11406-009-9218-0>.
- Yamada, Keiichi, and James K Kuchar. 2006. "Preliminary Study of Behavioral and Safety Effects of Driver Dependence on a Warning System in a Driving Simulator." *IEEE transactions on systems, man, and cybernetics-Part A: Systems and humans* 36 (3): 602–610.

- Yanco, Holly A, Munjal Desai, Jill L Drury, and Aaron Steinfeld. 2016. “Methods for Developing Trust Models for Intelligent Systems.” In *Robust Intelligence and Trust in Autonomous Systems*, 219–254. Springer.
- Zhang, Jiaje, and Donald A Norman. 1994. “Representations in Distributed Cognitive Tasks.” *Cognitive science* 18 (1): 87–122.
- Zhang, Zhimin, Huansheng Ning, Feifei Shi, Fadi Farha, Yang Xu, Jiabo Xu, Fan Zhang, and Kim-Kwang Raymond Choo. 2021. “Artificial intelligence in cyber security: research advances, challenges, and opportunities.” *Artificial Intelligence Review*: 1–25.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California