2021-09

EVALUATION OF THE SAFETY RISKS IN DEVELOPING AND IMPLEMENTING AUTOMATED BATTLE MANAGEMENT AIDS FOR AIR AND MISSILE DEFENSE

Cruz, Luis A.; Hoopes, Angela L.; Pappa, Ryane M.; Shilt, Savanna L.; Wuornos, Samuel I.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

SYSTEMS ENGINEERING CAPSTONE REPORT

EVALUATION OF THE SAFETY RISKS IN DEVELOPING AND IMPLEMENTING AUTOMATED BATTLE MANAGEMENT AIDS FOR AIR AND MISSILE DEFENSE

by

Luis A. Cruz, Angela L. Hoopes, Ryane M. Pappa,
Savanna L. Shilt, and Samuel I. Wuornos

September 2021

Advisor: Bonnie W. Johnson
Second Reader: Scot A. Miller

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2021	**3. REPORT TYPE AND DATES COVERED** Systems Engineering Capstone Report

4. TITLE AND SUBTITLE EVALUATION OF THE SAFETY RISKS IN DEVELOPING AND IMPLEMENTING AUTOMATED BATTLE MANAGEMENT AIDS FOR AIR AND MISSILE DEFENSE	**5. FUNDING NUMBERS**
6. AUTHOR(S) Luis A. Cruz, Angela L. Hoopes, Ryane M. Pappa, Savanna L. Shilt, and Samuel I. Wuornos	
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)

The modern battlefield is more complex than ever, and the technological advancement of weapons is accelerating. In order to win the next fight, faster response time to an adversary's actions is critical. Artificial intelligence (AI) has the potential to enable warfighters to outpace enemy decision cycles and reduce information overload, thus overcoming the "fog of war." When developing combat systems, reliability could be the difference between life and death. Therefore, it is of utmost importance that these weapon systems (especially novel systems such as AI) are developed with the highest standards of reliability and safety, long before they are introduced to the battlespace and entrusted to protect our nation's warfighters. This project utilizes a Systems Engineering approach to identify potential hazards and risks associated with AI and its role in the battlespace. Using an established Risk Management Framework (RMF), the team provides some mitigation strategies that developers must consider as they foster this technology for future use in U.S. weapon systems and processes.

14. SUBJECT TERMS artificial intelligence, AI, machine learning, safety risks, battle management aids, failure modes, air and missile defense	**15. NUMBER OF PAGES** 161
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

EVALUATION OF THE SAFETY RISKS IN DEVELOPING AND IMPLEMENTING AUTOMATED BATTLE MANAGEMENT AIDS FOR AIR AND MISSILE DEFENSE

Luis A. Cruz, Angela L. Hoopes, Ryane M. Pappa,
Savanna L. Shilt, and Maj Samuel I. Wuornos (USMC)

Submitted in partial fulfillment of the
requirements for the degrees of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING

and

MASTER OF SCIENCE IN ENGINEERING SYSTEMS

from the

NAVAL POSTGRADUATE SCHOOL
September 2021

Lead Editor: Samuel I. Wuornos

Reviewed by:

Bonnie W. Johnson	Scot A. Miller
Advisor	Second Reader

Accepted by:
Oleg A. Yakimenko
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The modern battlefield is more complex than ever, and the technological advancement of weapons is accelerating. In order to win the next fight, faster response time to an adversary's actions is critical. Artificial intelligence (AI) has the potential to enable warfighters to outpace enemy decision cycles and reduce information overload, thus overcoming the "fog of war." When developing combat systems, reliability could be the difference between life and death. Therefore, it is of utmost importance that these weapon systems (especially novel systems such as AI) are developed with the highest standards of reliability and safety, long before they are introduced to the battlespace and entrusted to protect our nation's warfighters. This project utilizes a Systems Engineering approach to identify potential hazards and risks associated with AI and its role in the battlespace. Using an established Risk Management Framework (RMF), the team provides some mitigation strategies that developers must consider as they foster this technology for future use in U.S. weapon systems and processes.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AAMD	Air and Missile Defense
ACS	Aegis Combat System
AI	Artificial Intelligence
BMA	Battle Management Aid
BMD	Ballistic Missile Defense
C2	Command and Control
C2BMC	Command Control Battle Management Communications
COA	Course of Action
COCOM	Combatant Command
CONOPS	Concept of Operations
CR	Concept Refinement
DOD	Department of Defense
DON	Department of the Navy
FOB	Forward Operating Base
HELIOS	High Energy Laser with Integrated Optical-dazzler and Surveillance
ICBM	Intercontinental Ballistic Missile
JP	Joint Publication
JTAMDO	Joint Theater Air and Missile Defense Organization
MDA	Missile Defense Agency
ML	Machine Learning
OV	Operational View
PM	Program Manager
RMF	Risk Management Framework
SDD	System Development and Demonstration
SOP	Standard Operating Procedure
TTP	Training Tactics and Procedures
UAV	Unmanned Aerial Vehicle
WF	Warfighter
WM	Working Memory

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The modern battlefield is more complex than ever, and the technological advancement of weapons is accelerating. In order to win the next fight, a faster response time to an adversary's actions is critical. Artificial intelligence (AI) has the potential to enable warfighters to outpace enemy decision cycles and reduce information overload, thus overcoming the "fog of war." Some examples of possible uses of AI include integrated battle management aids (BMAs) that help an operator decide, algorithms that predict future outcomes of engagements, and identification of friend-or-foe.

In order to employ AI effectively, developers must understand the benefits and risks associated with creating machines of war that can "think" like humans. Such risks are not limited to the technology but could also include the human dimension such as when warfighters distrust a computer to make decisions for them. Another example of potential risk is that the data that "trains" the AI could be faulty, old, or meaningless, rendering it ineffective. Additionally, the AI could "fail" by incorrectly choosing an action when faced with non-concurrence from another AI entity or BMA, resulting in threats impacting friendly targets.

When developing combat systems, reliability could be the difference between life and death. Therefore, it is of utmost importance that these weapon systems (especially novel systems such as AI) are developed with the highest standards of reliability and safety, long before they are introduced to the battlespace and entrusted to protect our nation's warfighters. This project utilized a systems engineering approach to identify potential hazards and risks associated with artificial intelligence and its role in the battlespace. Using an established Risk Management Framework (RMF), the team provides some mitigation strategies that developers must consider as they foster this technology for future use in U.S. weapon systems and processes.

The team also employed systems engineering to conduct the project analysis. First, they oriented on the problem and defined requirements. To accomplish this, the team learned what exactly AI and machine learning (ML) are by conducting an extensive

literature review of prior works on the subject. This enabled the team to develop system architecture diagrams to understand potential system structure and hierarchy. The team then drew on personal knowledge from within its membership (such as two members who work for the Missile Defense Agency and one Active Duty Marine Officer) to develop use case scenarios for potential employment of AI in the battlespace. Using Innoslate to develop artifacts, the team then conducted a safety analysis from these use cases to identify hazards and failure modes. These hazards and failure modes were analyzed using the RMF from the *National Institute of Standards and Technology Special Publication 800-37 Revision 2*. This enabled the team to develop mitigation strategies for the identified hazards.

As stated above, the team developed three use cases: (1) a ballistic missile defense scenario, (2) a ship under attack from a swarm of unmanned aerial vehicles (UAV), and (3) a scenario in which theater-level and strategic-level AI systems produce contradictory recommendations. The team chose these scenarios based on the level of impact they could make on the nation (such as with a ballistic missile armed with a nuclear warhead), their likelihood (such as with a high-payoff target like a large naval vessel), and the future of warfare shifting to an expeditionary nature (such as forward operating bases (FOBs) and expeditionary advanced bases). Failure modes and mitigation strategies were extensive for each scenario (as well as for common system hazards for computer assets). By identifying these failure modes and mitigation strategies, the team provides a baseline for future planning against other possibilities and scenarios.

Scenario 1's ballistic missile defense situation highlights Warfighter Mistrust. In this scenario, warfighters react to an incoming ballistic missile based on their own concept of operations, instead of what the AI recommends. The hazards associated with this mistrust include ineffective response time, ineffective countermeasures, incorrect lethal object selection, and improper location/timing of where the countermeasure will impact. Scenario 2's ship self-defense situation focused on Training Data for the AI's development. The team identified such hazards as misidentification and ineffective responses, along with failure modes associated with each. Scenario 3's primary hazards were derived from the principal mishap of a successful enemy attack on a friendly FOB. The hazards that allow

this mishap to occur are the hostile threat not being neutralized, and whether or not it is not engaged at all.

The team developed mitigation strategies for each of these scenarios. Scenario 1's prime strategies were to establish time standards for the AI to adhere to in the decision-making process, and for user concept of operations (CONOPS) to be updated regularly, as well as in the pre-deployment phase. For Scenario 2, proper programming techniques in pre-deployment, regular (monthly) updates to the training data, and utilization of back-up data would prevent misidentification and ineffective responses. Scenario 3's hazards can be mitigated by proper programming in the pre-deployment phase with Joint forces' input.

In the end, the team recommends that further study take place on how to implement AI/ML at a Tactical and operational level, that AI/ML are used to gather performance data on new or existing threats, that the DOD directs how verification and validation will be managed for systems that will use AI/ML, and that a service level and DOD level reliability study for AI/ML BMAs is conducted. By the end of this report, the reader should have a better understanding of how AI/ML can benefit the warfighter, and what precautions must be taken to ensure it is developed as safely as possible.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

The team would like to thank Dr. Bonnie Johnson and Scot Miller of the Naval Postgraduate School's Systems Engineering and Information Sciences Departments, respectively. Thank you for your guidance as we started with nothing but a topic, and you helped us grow it into a full-fledged capstone.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Lieutenant Junior Grade Smithers sat comfortably at his desk in the Combat Information Center of his ship, an Arleigh Burke class Destroyer. The room was affectionately referred to as "Combat." He had just earned his Surface Warfare pin, a major milestone in a naval officer's career. Now that he was the youngest fully qualified officer on the ship, he had been assigned to the graveyard watch, but he did not mind; he enjoyed the peace and quiet.

A noisy alarm and flashing lights from his screen interrupted his reverie. His designated Officer of the Watch master screen automatically cycled through three windows before it settled on an image showing the entire western Pacific.

Figure 1. Officer of the Watch Screen. Source: NOAA.

He recognized a cluster of blips as his own carrier strike group, but he did not know what to make of everything else. His heart raced as he tried to determine if this was a drill or not; there was not one scheduled for this time of night. He did not notice the hum of the ship's capacitors charging.

Figure 2. Sailor at Watch Station. Source: MC3 Cosmo Walrath/U.S. Navy.

The window cycled itself again, highlighting a red icon, moving much faster than any blip should be moving, faster than any jet he was aware of. The highlighted blip was accompanied by a textbox identifying it as:

Inbound Missile.

Target: USS THEODORE ROOSEVELT (95%)

Engage Target?

"That can't be right," Smithers said aloud as he picked up the phone to the Captain's quarters. He dialed the number and looked back at the screen. The blip was now uncomfortably close to the strike group's cluster. Much too close. "Wait," Smithers said aloud.

"Don't tell me to wait Smithers. What's going on?!" the Captain's disembodied voice barked at him through the phone. The lights dimmed and thunder clapped from the upper decks. Smithers's screen now showed a new message:

"Target destroyed - approx. 1.1NM from USS THEODORE ROOSEVELT. Follow-on attack imminent. Recharging HELIOS."

"Uhhh, sir, I think 'George' just shot down a cruise missile..."

This fictional scenario illustrates one potential use of artificial intelligence (AI) in combat. The reader may have noted that Lieutenant Junior Grade Smithers hesitated to act upon the initial notification of the impending danger. The AI (referred to as 'George' in this vignette) anticipated this hesitation, and automatically powered up the ship's onboard missile defense system (in this case, a LASER system known as HELIOS). The AI also utilized Smithers's Officer of the Watch screen to present a simplified decision space so as not to overwhelm him. When Smithers exhibited a natural skepticism of the information presented to him, the AI made the choice to shoot down the missile when it crossed a pre-established threshold.

A. BACKGROUND

The concepts of automation and AI have been around for many years. Gregory Allen (2020) states, "Though many AI technologies are old, there have been legitimate technological breakthroughs over the past ten years that have greatly increased the diversity of applications where AI is practical, powerful, and useful." Machine learning (ML) is a subset of the field of AI and has been the focus of many research efforts recently. Figure 3 illustrates the connection between automation, AI and ML.

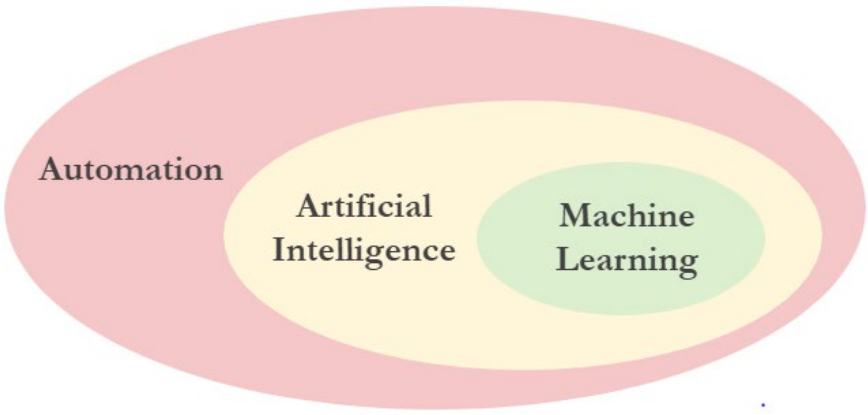

Figure 3. Venn Diagram of Automation, AI and ML. Source: Johnson (2021).

Artificial intelligence/machine learning offers the potential of improving warfighters' situational awareness of the battlespace and improving the process and speed of tactical decision-making in time-critical and complex threat situations. The benefits will not come without the potential for safety risks during implementation of AI and ML. Figure 4 depicts some of the safety risks associated with the use of AI and ML in battle management aids. Automated systems are vulnerable to cyber-attacks, operators may experience trust or interaction issues, and ML systems in particular, are susceptible to providing skewed or biased outcomes.

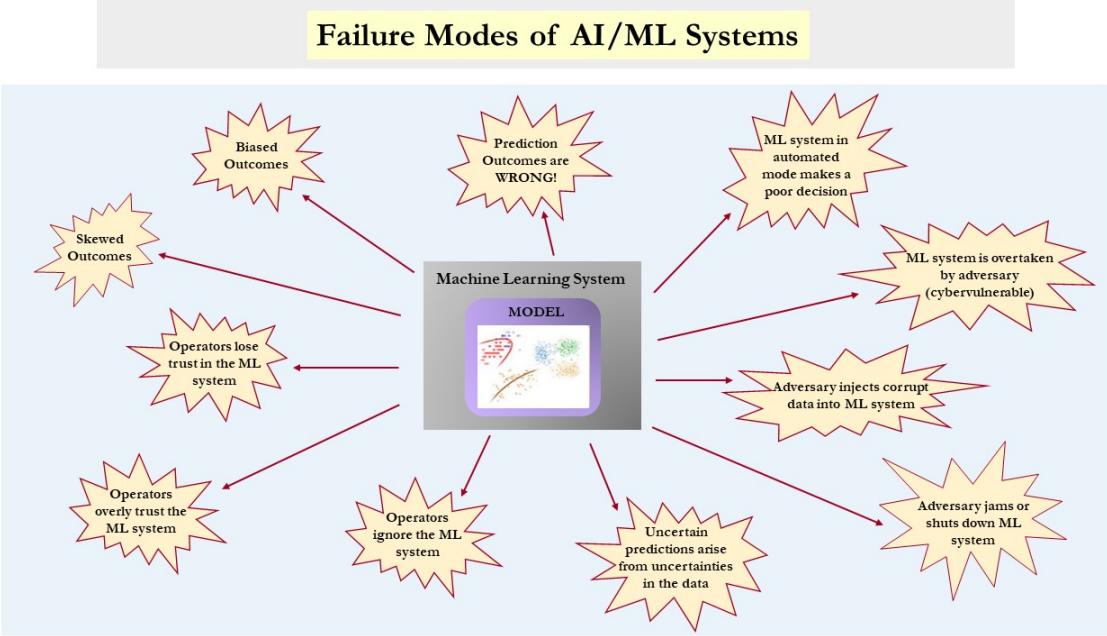

Figure 4. Examples of Failure Modes of AI/ML Systems. Source: Johnson (2021)

Artificial intelligence is becoming increasingly more attractive to the DOD as a capability with wide-ranging applications. According to the 2018 DOD Strategy on AI, "The costs of not implementing this strategy are clear. Failure to adopt AI will result in legacy systems irrelevant to the defense of our people, eroding cohesion among allies and partners, reduced access to markets that will contribute to a decline in our prosperity and standard of living, and growing challenges to societies that have been built upon individual freedoms" (DoD 2018). In particular, the air and missile defense (AAMD) mission area is of particular interest given the complexity in ballistic missile defense, cruise missile defense, hypersonic missile defense, and air defense. Multiple defense systems exist to defeat threats at various stages of flight controlled by human warfighters. In some instances, these human warfighters become overwhelmed when the decision space becomes complex due to time constraints, information challenges (too much, too little, or too faulty), or threat challenges (multiple and/or diverse AAMD threats). Including an automated decision aid to assist the warfighter, or even take on the role of the decision-maker, is a domain space being explored in many parts of DOD (DoD 2018).

The following two operational views (OVs) depict the use of battle management aids with AI/ML at a strategic level (Figure 5) and a regional level (Figure 6), along with embedded risk charts that identify some of the safety risks that need to be investigated.

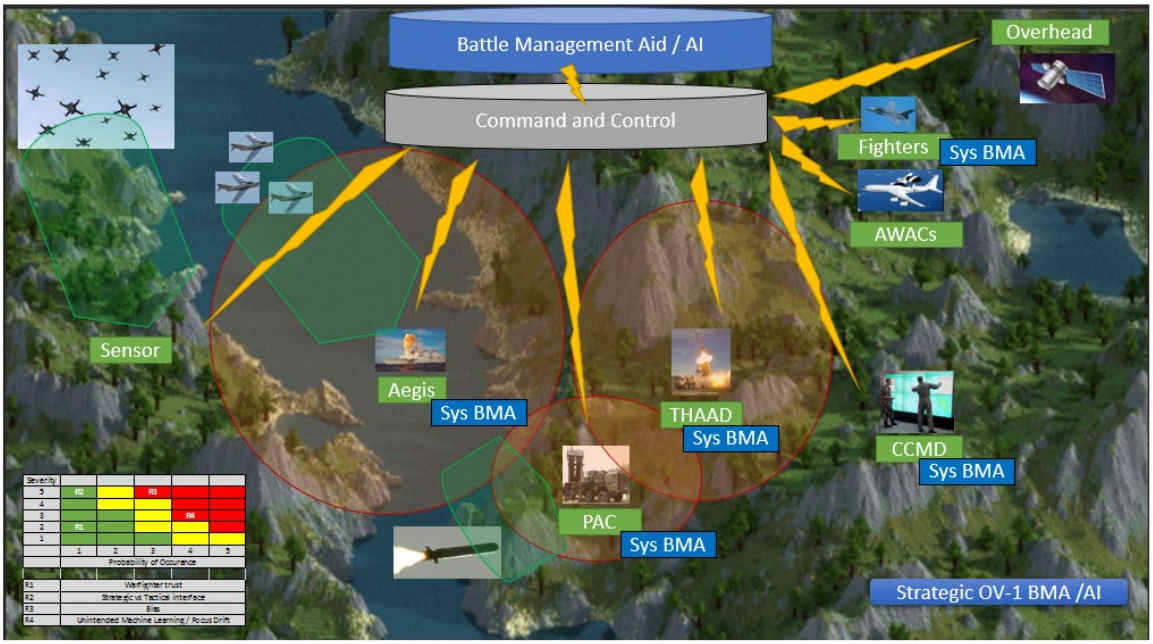

Figure 5. Strategic Level OV-1 – Safety in Automated Battle Management Aids

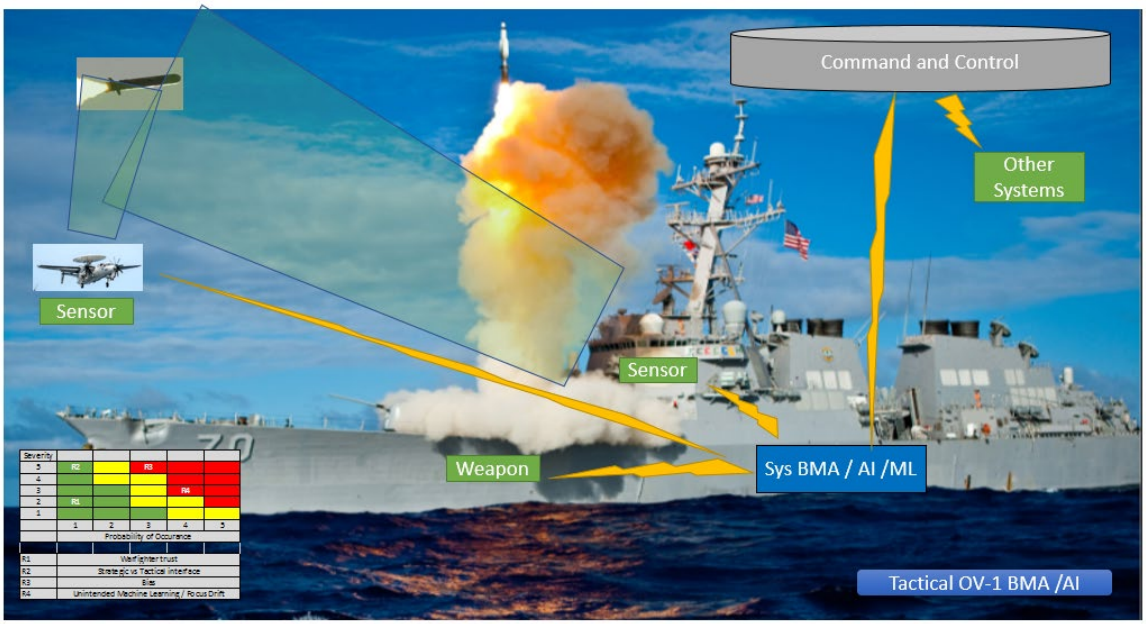

Figure 6. Regional Level OV-1 – Safety in Automated Battle Management Aids

Given the high likelihood that AI and ML will be integrated into command and control, battle management aids, and the weapon systems themselves, this capstone project explored the potential hazards in introducing AI and ML capabilities as an automated battle management aid (BMA) for the AAMD mission.

B. PROBLEM STATEMENT

The advancement of technology has increased the speed of warfare requiring faster reaction times and human decision making. The Department of Defense (2017) has acknowledged the necessity to acquire tactical decision aids for the purpose of alleviating the stress of battlefield decision making for commanders and warfighters. The use of automated methods, including AI and ML, in BMAs can help to meet diverse mission needs as well as assist with the transition from planning to execution (Department of Defense 2017). However, the use of AI and ML in future BMAs introduces safety risks and new failure modes due to the non-deterministic and evolving nature of AI systems, the complex human-machine interactions, and challenges related to the development and operation of a learning system.

C. PROJECT OBJECTIVES

The objective of the capstone project was to study the safety risks related to the development and implementation of future BMAs leveraging AI and ML for the AAMD mission. Specifically, this study addressed the following questions:

- What are the safety risks related to the deployment of AI systems that support future automated tactical decision and mission planning aids?
- What are the possible consequences of safety related problems in AI systems used in tactical decision making?

D. STAKEHOLDERS

The team identified key stakeholders and assessed their needs as shown in Table 1. The stakeholders include organizations and end users who will benefit from this study. End users (warfighters), in particular, will benefit from the implementation of successful and safe BMAs that leverage AI and ML capabilities. Program managers and engineers can incorporate the results of this study into system requirements and designs for safe AI/ML BMAs for the AAMD mission.

Table 1. Key Stakeholders

Key Stakeholders	Needs
End User (Servicemen)	Trust, safety enhanced tactical decisions and ease of use
Department of Defense (DoD)	Enhanced reaction time and more educated war decisions which will improve the safety of the war fighter and DoD assets
Department of Navy (DoN)	
Missile Defense Agency (MDA)	
DoD Cyber	Cyber security, ensuring our assets and information are safe and not compromised
Secondary Stakeholders	**Needs**
Battle Management Aid (BMA) / Artificial Intelligence (AI)	
Program Managers (PMs)	Proper coordination with Key Stakeholders to determine their needs and ensure needs are met
Chief Engineers	Proper study and conceptual design to be able to build and implement BMAs with AI
System Engineers	
Product Engineers	
Contractors	

E. TEAM ORGANIZATION

The capstone team consisted of the following NPS systems engineering students: Angela Hoopes, Luis Cruz, Ryane Pappa, Savanna Shilt, and Samuel Wuornos. Table 2 introduces the team's roles and their respective organizations.

Table 2. Project Team Membership

Team Member	Role	Organization
Angela Hoopes	Team Leader	Systems Assessment Team Lead Engineer NH-04 0801, Missile Defense Agency - Aegis BMD Program Office - Engineering Directorate
Luiz Cruz	Development and Integration Lead	Director for Test, Israeli Cooperative Program Office, Missile Defense Agency
Ryane Pappa	Engineering Lead	General Engineer Team Lead DB-03 0801, Systems Engineering Directorate, U.S. Army Combat Capabilities and Development Command Armaments Center (DEVCOM-AC)
Savanna Shilt	Lead Analyst	Computer Scientist NH-03 1550, United States Army Information Systems Engineering Command (USAISEC)
Major Wuornos, Samuel	Lead Editor	Aircraft Maintenance Officer and Pilot, Marine Heavy Helicopter Squadron 466, Marine Aircraft Group 16, 3rd Marine Air Wing, United States Marine Corps

The Team Organizational Chart in Figure 7 describes the high-level organizational structure for Team A.ctual I.ntelligence that includes the roles of Capstone Advisor, Team Lead, Second Reader, Modeling Lead, Engineering Lead, Lead Editor, and Lead Analyst.

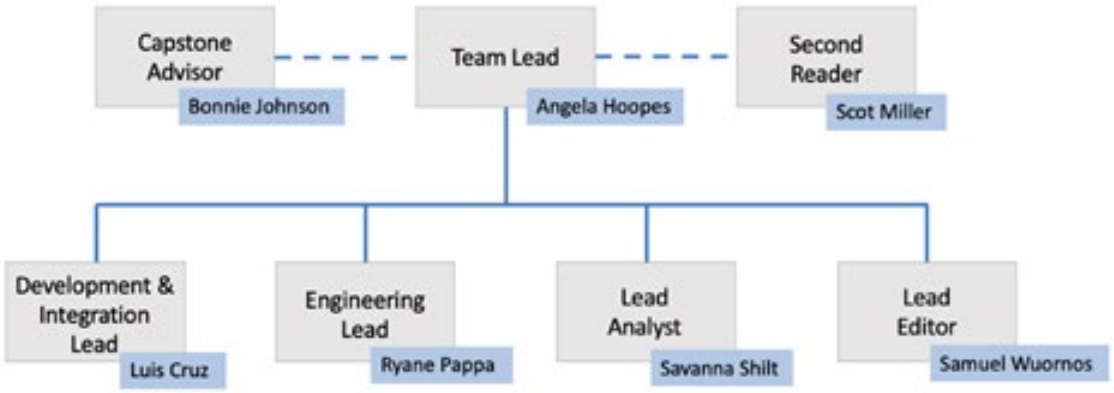

Figure 7. Team Organization

The responsibilities given to each role presented in the organizational chart were established based on the key studies and activities that would be completed throughout the course of the Capstone Project. Table 3 identifies each team member and their roles and responsibilities.

Table 3. Project Team Membership

Team Member	Role	Responsibility
Bonnie Johnson	Capstone Advisor	Supervises the project execution to include approval of the Capstone project and final report.
Scot Miller	Second Reader	Provides engineering experience and background to help guide the Capstone research and discussions.
Angela Hoopes	Team Lead	Leads the project team in conducting and completing the Capstone project. Organizes and delegates project task as needed and serves as the primary liaison to the Capstone Advisor.
Luis Cruz	Development and Integration Lead	Leads the translation of requirements analysis and system architectures into a solution that can be deployed across the DoD.
Ryane Pappa	Engineering Lead	Leads engineering analysis and requirements analysis. Develops future AI and ML BMA system alternatives.
Savanna Shilt	Lead Analyst	Leader of project analysis including the risk management process.
Samuel Wuornos	Lead Editor	Responsible for ensuring all project deliverables meet format and style standards prior to formal delivery.

F. PROJECT APPROACH

The team utilized a systems engineering approach to conduct the analysis for this project. First, they oriented on the problem and defined requirements. To accomplish this, the team focused on learning what exactly AI and ML are by conducting an extensive literature review of prior works on the subject. This enabled the team to develop system architecture diagrams to help them understand potential system structure and hierarchy. The team then drew on personal knowledge from within its membership (such as two members who work for the Missile Defense Agency and one Active-Duty Marine Officer) to develop use case scenarios for potential employment of AI in the battlespace. Using Innoslate to develop artifacts, the team then conducted a safety analysis from these use cases to identify hazards and failure modes. These hazards and failure modes were then analyzed using the Risk Management Framework (RMF) from the *National Institute of Standards and Technology Special Publication 800-37 Revision 2*. This enabled the team to develop mitigation strategies for the identified hazards.

G. CAPSTONE REPORT OVERVIEW

Chapter I provided an introduction to and background for the project. It presented the problem statement, project objectives, stakeholder description, team organization, and project approach.

Chapter II provides a review of previous works that were researched by the team. These works offer key background information on machine learning, artificial intelligence, and warfighter decision-making. This chapter describes why these works are relevant to this project.

Chapter III covers the critical analysis of three use case scenarios involving AI/ML in missile defense. The use cases include Ballistic Missile Defense, Naval Warship Self Defense, and Strategic vs Theater Bias. The chapter discusses identified failure modes and hazards in detail, providing a baseline for risk assessment.

Chapter IV builds on the analysis from Chapter III and presents an in-depth risk analysis of the identified failure modes and hazards of each use case. The team uses this risk analysis to provide mitigation strategies for future developers to consider.

Chapter V addresses conclusions wrought from the previous chapters and discusses potential ways forward in the development/procurement of AI/ML with regards to missile defense and future combat systems and processes.

II. REVIEW OF PRIOR WORKS

A literature review was conducted to understand the various subjects related to the problem statement. The team reviewed articles and papers ranging from the AAMD mission, what is AI/ML and why is it needed, and challenges and safety risks with the introduction of AI/ML into a system of systems. The information provided in this chapter helps to align the reader with how the team framed the problem.

A. WHAT IS AI/ML?

The study of AI and ML is broad in scope, but it is best to start with basic definitions. Unfortunately, there are many definitions for both terms out there. A professor at Dartmouth College in 1955, John McCarthy, is known as having first defined artificial intelligence. McCarthy defined artificial intelligence as "The science and engineering of making intelligent machines" (2007, 2). Bernard Marr is a futurist and states "The focus of artificial intelligence shifts depending on the entity that provides the definition" (2021, 1). Marr provides six definitions of AI that all vary slightly. Some are provided from various dictionaries while others are based on companies that invest in AI and the objectives the company wishes to achieve. The DOD Artificial Intelligence Strategy states "AI refers to the ability of machines to perform tasks that normally require human intelligence – for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action." (2018, 5).

DeepAI is an artificial intelligence community interested in technology development for the future. Their website, https://deepai.org, contains extensive amounts of research, news, guides and information on the field of AI. They define machine learning as "a field of computer science that aims to teach computers how to learn and act without being explicitly programmed" (2021, 1). The Berkeley School of Information describes the idea of machine learning as "using statistical learning and optimization methods that let computers analyze datasets and identify patterns" (Tamir 2021, 1). Machine learning uses algorithms to process large amounts of data to arrive at the next step or next decision that

needs to be made. The algorithm must continuously learn from the data it analyzes and constantly improve its output without requiring human interaction.

The concepts of automation and artificial intelligence are often used interchangeably even though they are different. Wang and Siau (2019) state that "automation frees humans from time-consuming and repetitive tasks" (2019, 63). Ideas about automation typically revolve around manufacturing processes that involve completing the same tasks repeatedly. Automation is also seen in software with the use of programmed rules to complete repetitive tasks. This is where some of the confusion lies with the two terms. Automation can only perform repetitive tasks using a pre-programmed ruleset whereas AI is able to learn from patterns and apply what it has learned to new data, essentially mimicking human intelligence.

There have been many improvements and use cases for AI/ML, especially over the last twenty years, as shown in Figure 8. Allen (2020) identifies four key factors responsible for improvements in ML performance as shown in Figure 9. These four factors have had a large impact on various use cases for machine learning that were once either almost impossible or too expensive.

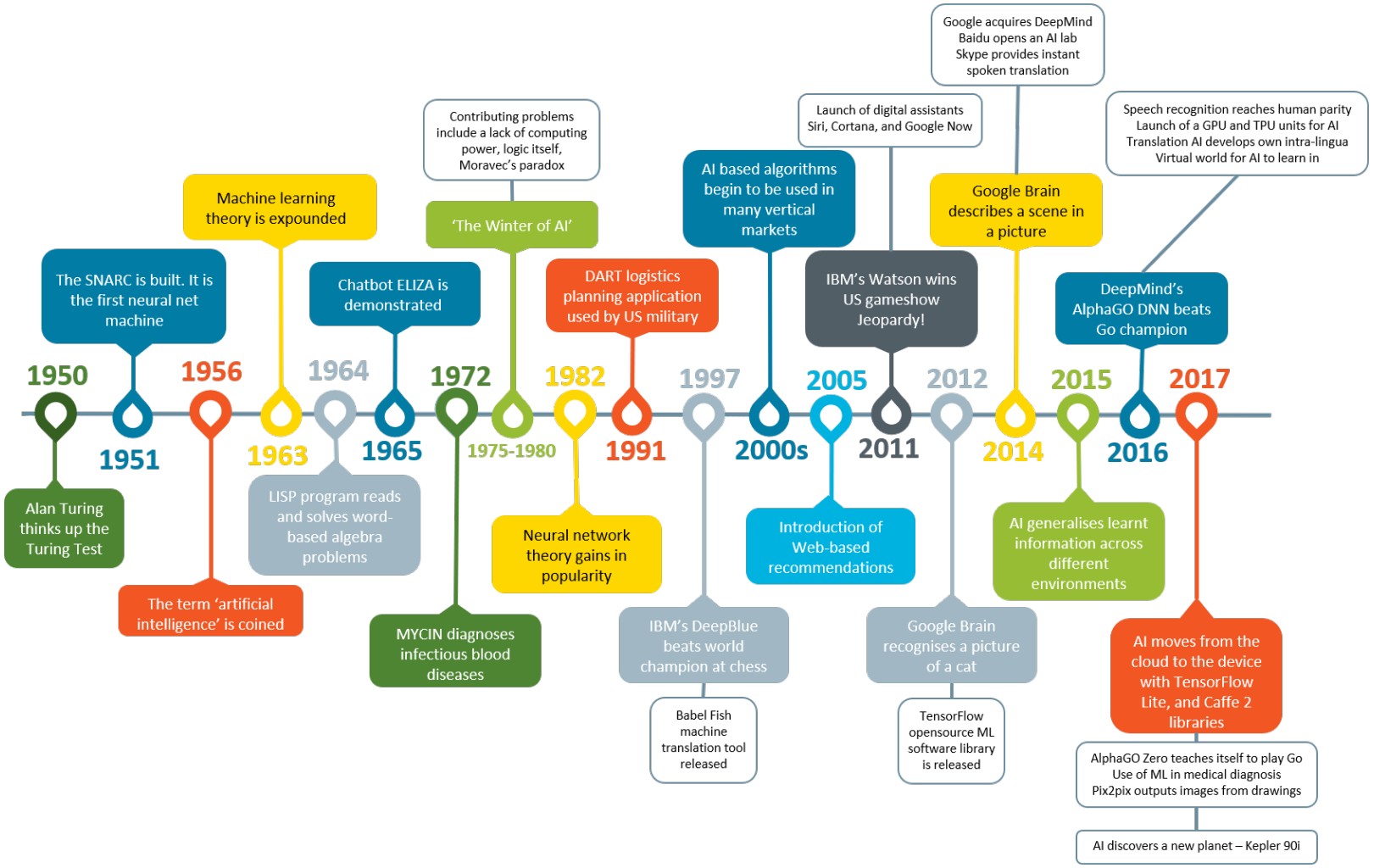

Figure 8. AI/ML Timeline. Source: SeekPNG (2019).

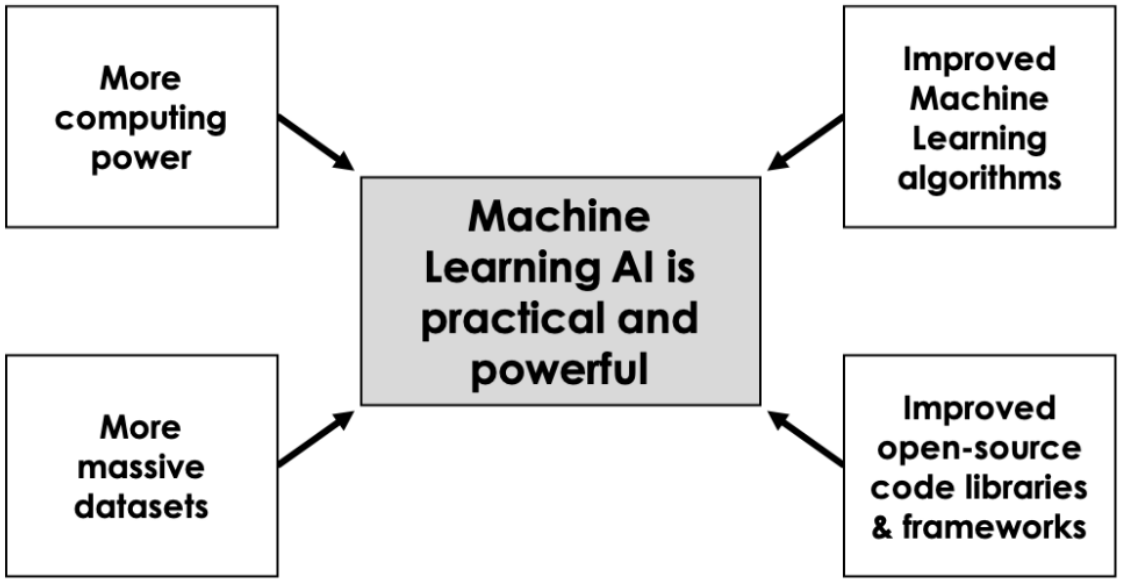

Figure 9. Four Key Factors of Machine Learning. Source: Allen (2020).

There are several types of acknowledged machine learning in various works. The more common instances are supervised, unsupervised and reinforcement learning. In supervised learning the data inputs are labeled according to respective data outputs prior to the algorithm processing any training data. Data needs to be properly labeled to ensure the most accurate system performance. Unsupervised learning is exactly the opposite since the data are not pre-labeled. Algorithms that employ unsupervised techniques can extract various categories or features from the data. This can be good or bad, as the data may extract features that were unanticipated by someone looking at the outputs. Unsupervised learning is a good method to initially appraise data when the relationships between data are not known or if the data set is too large to determine relationships. Reinforcement learning is a method that enables algorithms to learn from observations made in the environment. The algorithm takes an action which the environment responds to. The algorithm learns from that new environment and can take another action. This type of learning is often seen in digital games such as chess and various card games.

Wang and Siau (2019) discuss AI, ML and the benefits those capabilities have had across various use cases. They state, “AI has enormous potential in business,

manufacturing, healthcare, education, military, and many other areas" (Wang and Siau 2019, 63). They provide many real-world applications of AI shown in Table 4. There are many benefits to the use of AI but there are also many risks to using AI. Wang and Siau identify some of these risks, which are highlighted in Section C of this chapter.

Table 4. AI Use Cases and Their Impacts. Source: Wang and Siau (2019).

AI Use Case	**Impact on Real World**
Self-driving vehicles	Fewer traffic accidents
Education	Attending class from home
Human Resources	Efficient application review/processing
Cybersecurity	Threat detection and response
Home	Automation of lights, thermostats, etc.
Health Care	Patient risk assessments
Finance	Fraud monitoring and identity theft

B. HOW AI/ML COULD CHANGE THE BATTLEFIELD

For this project, it was essential to review how the battlefield of today is changing and how AI/ML could be part of that change. An increasing number of sensor and weapon systems interact to create a common tactical picture for accurate situational awareness for the warfighters and commanders. This common tactical picture is created via massive amounts of data fusing together to output the important, and often, time-sensitive data. Figure 10 depicts many high-speed, simultaneous engagements and complex sensor networks which could quickly and easily overwhelm a warfighter.

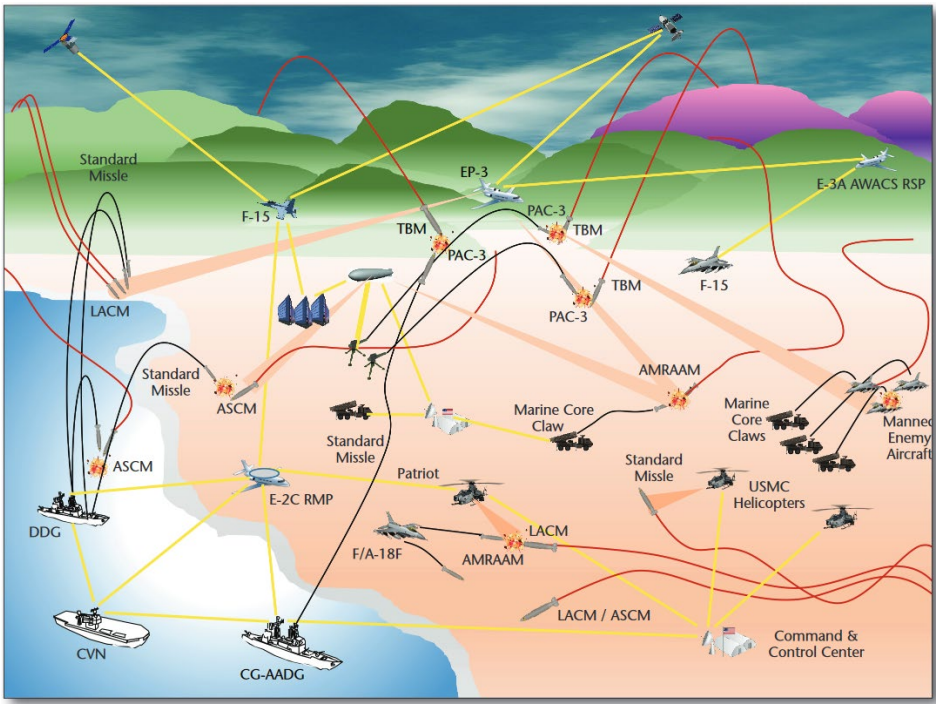

Figure 10. Battlefield Complexity. Source: Johnson (2019).

In a paper by Soller and Morrison (2008), research was done to assess battle managers' performance against specific automated tasks by automated battle management aids. They discussed studies that were made in an "operator-in-the-loop" environment to determine how air and missile defense battle managers degrade with workload and how automation assisted the warfighter. The paper also acknowledged research done by Kaempf, Wolf, and Miller (1993) that studied decision making processes on Aegis cruisers and found that "the most difficult part of an operator's task involved assessing the situation and obtaining the information needed to maintain good situational awareness vice engaging a threat" (Soller and Morrison 2008, 17). Engagements were binned into four basic areas: (1) acquisition of information, (2) representation and display of the information, (3) decision-making, and (4) implementation. These four areas can leverage battle management aids as tools to assist the warfighter. Given the complex tactical environments the warfighters are operating in today, AI/ML could help warfighters in all four of these areas by quickly gathering and processing the incoming data, displaying the important information needed for a decision then executing the result of that decision.

Johnson and Treadway state similar thoughts in their paper *Artificial Intelligence – an Enabler of Naval Tactical Decision Superiority* (2019). They state, "AI enables BMAs for improving combat identification, identifying and assessing tactical courses of action, coordinating distributed warfare resources, and incorporating predictive wargaming into tactical decisions" (Johnson 2019, 1). The use of AI/ML in these areas would be extremely powerful for the warfighters giving them the leverage they need on today's battlefield. Two such programs were described in Grooms (2019) NPS Capstone titled, *Artificial Intelligence Applications for Automated Battle Management Aids in Future Military Endeavors*. DARPA's decision battle management (DBM) program uses AI to improve situational awareness and the BAE Company was able to improve mission effectiveness using semi-autonomous software. The use of automated BMAs in various military missions today will improve the warfighter's situational awareness and help to improve the common operating picture.

Intelligence, whether human or machine, needs to use things it knows and apply that knowledge to learn and understand things it does not know. Johnson (2021) references a quote from Donald Rumsfeld regarding unknown unknowns to understand the difficulties with combat identification as shown in Figure 11. In Figure 12, Johnson illustrates where the use of various AI methods can address what is unknown. So much data is produced by today's weapon systems, especially with a system of systems, that it is nearly impossible for humans to take that data, understand it all and provide decisions on a path forward. This is where AI/ML could be invaluable.

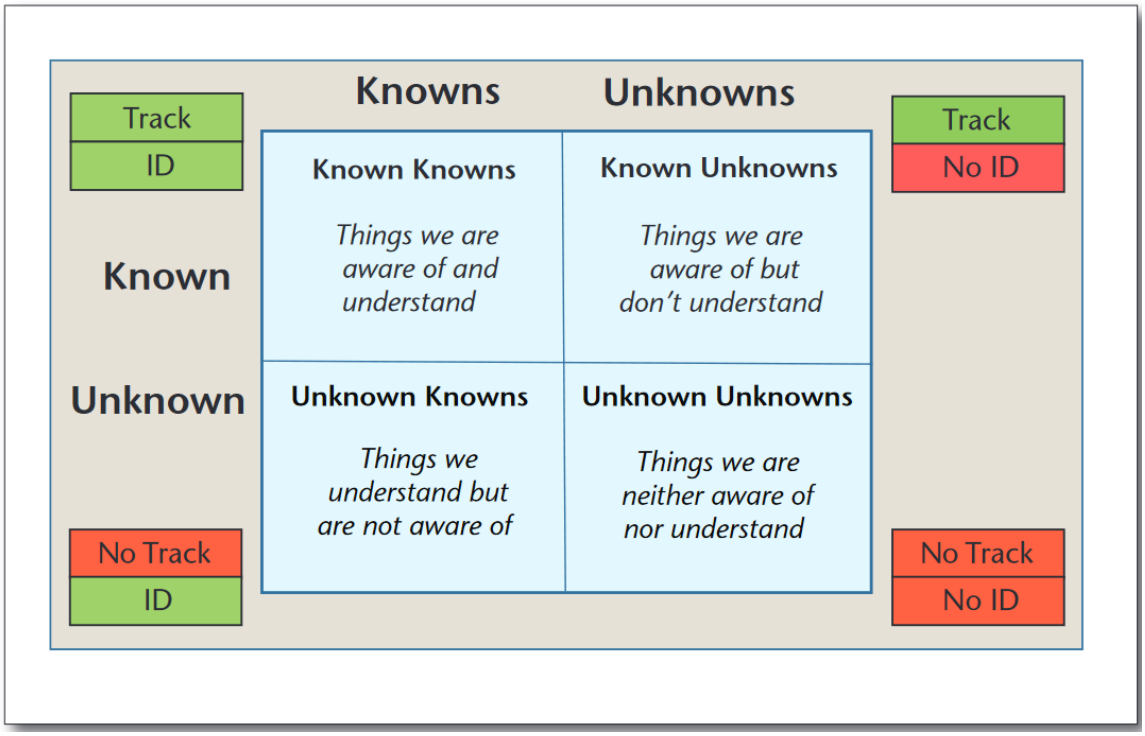

Figure 11. Knowns and Unknowns. Source: Johnson (2019).

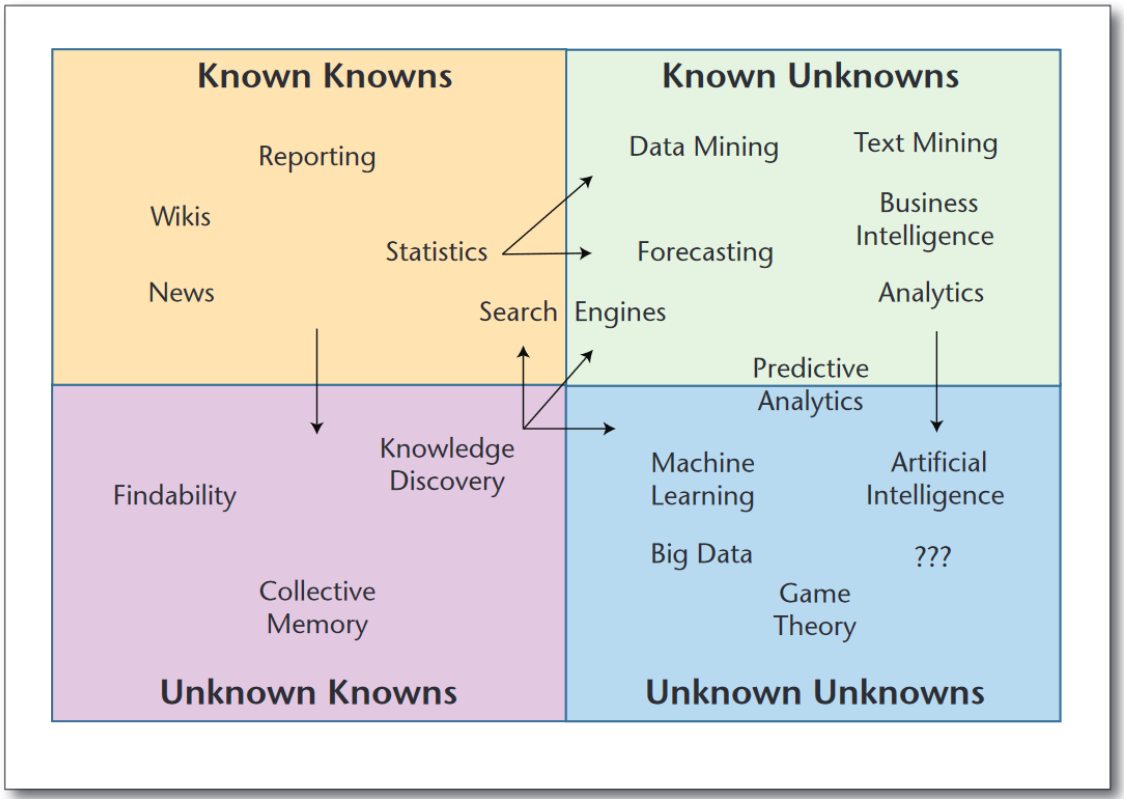

Figure 12. AI Methods for the Knowns and Unknowns. Source: Johnson (2019).

Grooms (2019) also studied the issue of combat identification and how AI/ML could help in this area. He interviewed five combat identification experts who possessed high proficiency across various tactical environments on where AI could help. Conclusions from his NPS Capstone concluded that the warfighters need increased situational awareness, testing of BMAs with warfighters in the loop, efficiency, and a user-friendly system, in both the ship's equipment and reports produced onboard. Using AI/ML in these areas would provide great benefits to warfighters, while including their recommendations during system design and development would ensure the end product's usefullness.

Studies have also been completed in the area of human vs. AI performance. In another NPS Capstone by Jones et al. (2020), they analyzed scenarios involving a single threat. The scenarios varied the levels of stress and mixes of AI. Scenarios ranged from human-in-the-loop with a low stress scenario to a high stress scenario with full automation

and use of AI. Analysis showed that efficiencies would be achieved by incorporating higher levels of automation and machine learning. Jones et al. acknowledge that gains would not be realized in the near term, however, as it would require more refined AI methods, a large amount of training data to help the system learn and building of operator trust in the system once more automation is applied.

Another area where AI/ML implementation would be useful is in the Operations Planning process. McKendrick (2017) refers to various services planning processes along with the Joint Operations Planning Process used by the United States and the Comprehensive Operations Planning Process used by NATO. She talks about how the capabilities to create a plan, see the plan play out, and adjust plans in a short time period would be of great benefit to those in the decision making process. Artificial intelligence/ machine learning could also help to assign assets to tasks and enable easy adaptation as conditions changed. Automating task monitoring for key indicators during battle would assist commanders in maintaining a more accurate picture of mission progress.

Using predictive analytics as a capability to support BMA automation is another area that could give warfighters the leverage they need to increase their battlefield effectiveness. Johnson illustrates her thoughts on a predictive analytics capability using the conceptual framework shown in Figure 13. The idea behind this capability is that various courses of action are developed based on models of Blue Force and SA knowledge. Each course of action is evaluated using a Red Force model to assess first and second order effects on the enemy. Those results are then analyzed to predict the next state of action for the Blue Force.

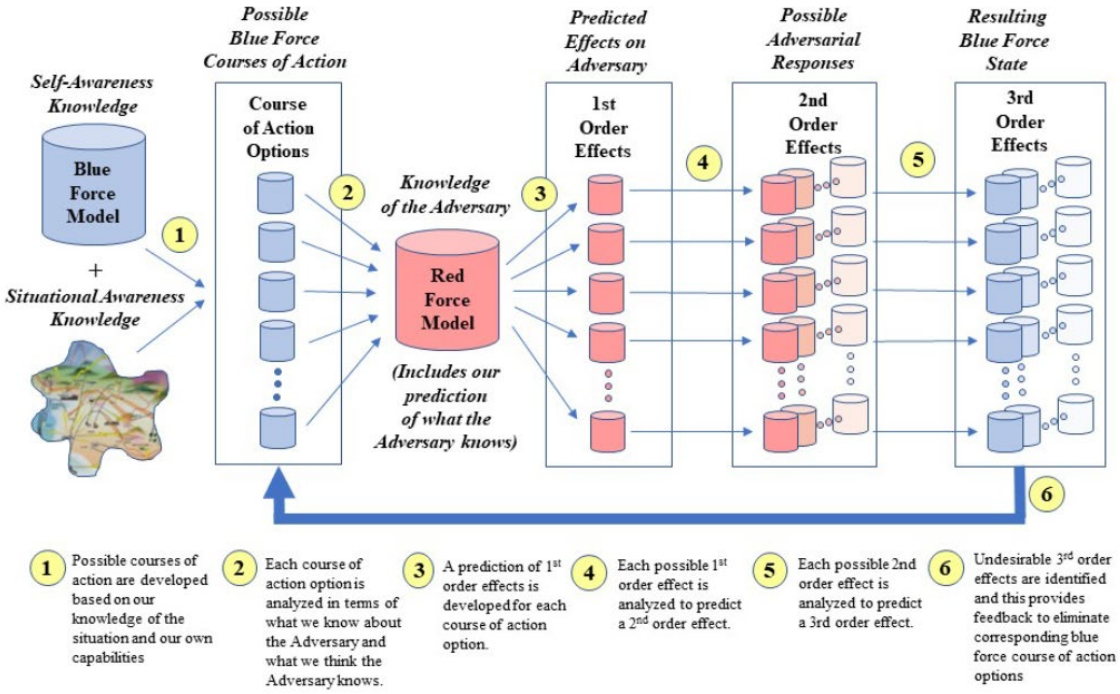

Figure 13. Conceptual Framework for Predictive Analytics Capability. Source: Johnson (2020).

C. CHALLENGES OF AI/ML SYSTEMS

Introducing AI/ML into battle management aids to assist the warfighter in the decision-making process will be beneficial but will also present many challenges. The research done by this project team identified many challenges. In a paper by Dr. Bonnie Johnson (2021), she identifies four unique challenges. First, is that today's warfare environment is very complex. The second challenge is that a large amount of training data is needed for the system. The third challenge she identifies are the new methods of systems engineering that will be required to ensure a safe system. Last, is remembering that our adversaries are also advancing their AI/ML capabilities.

These four challenges encompassed much of the team's research. The sheer number of weapon systems, sensors and targets provides insurmountable amounts of data to battle management aids with requirements for high update rates, which can overwhelm warfighters and combatant commanders, shrinking their decision space. Often, a human

operator is still in the decision-making loop, where the challenges include operator error and lack of warfighter trust in the decision produced by the AI/ML BMA.

The large amount of training data needed to help the systems learn behaviors or outcomes must be relevant to the algorithms being used and of high quality. According to a discussion with the Director for Machine Learning at Chess.com, it is all about the training data available (Terwillinger 2021). Any program using machine learning to better arrive at a decision is only as good as the information provided to the program. Therein lies the risk to consuming bad training data that would produce less-than-favorable results in a game of chess, engagement of ballistic missiles, or deciding to yield to incoming traffic. Johnson (2020) illustrates the decisions that program managers need to make when developing an AI/ML enabled system in Figure 14.

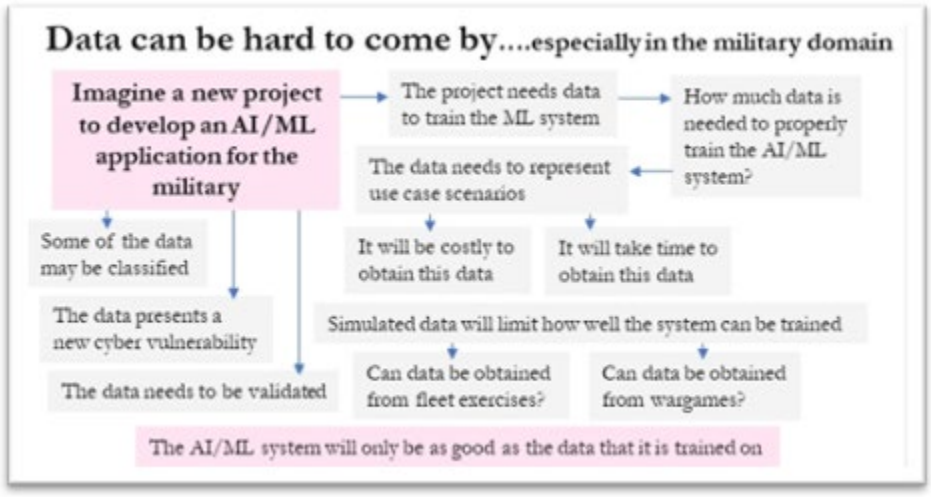

Figure 14. Development of Datasets for Artificial Intelligence and Machine Learning System Training. Source: Johnson (2021).

All these decisions impact cost, schedule, or performance of the AI/ML system. The integration of AI/ML into system processes will require changes to how the system is engineered, especially with respect to safety. The AI/ML system will constantly learn and adapt to changes in data that are provided. A system engineer will have to allow for the system behavior to evolve while ensuring it remains safe and trustworthy for the

warfighters. Also, because the system learns and adapts, the system engineer must also understand that system failures will be unlike other past failures, and therefore will require more explanations.

Johnson (2021) describes how our adversaries present challenges to the AI/ML system. Much of the research surrounding adversarial challenges stem from cyber-attacks and the ability for potential insider threats to corrupt training data fed to the system. The ability to outpace our adversary's development of AI/ML capabilities and protect our systems from attack will be more important than ever to give the warfighters the best system possible.

Wang and Siau (2019) also present their understanding of challenges and issues associated with AI systems. They binned these challenges into four categories as shown in Figure 15. Many of these challenges are not safety related but they still present an issue for those working to develop AI systems and provide stakeholders considerations. This project will address these issues as the three AAMD scenarios are investigated.

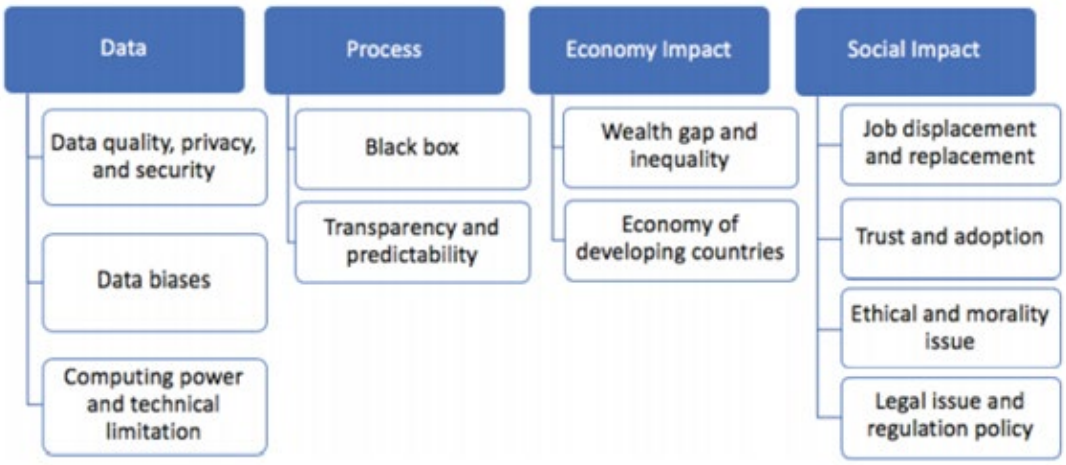

Figure 15. Challenges and Issues. Source: Wang and Siau (2019).

There has been significant progress made in identifying various failure modes associated with AI/ML systems. Johnson (2021) created a table of possible AI/ML failure modes as shown in Table 5. These failure modes can be seen in any system that

incorporates AI/ML. Miller and Nagy (2017) refer to the Naval Ordnance System Safety Activity (NOSSA) and various root causes for AI system failures that are seen. Table 6 depicts AI System Failure root cause examples.

Table 5. Examples of AI Failure Modes. Source: Johnson (2021c).

Failure Category	Failure Mode Examples
System Produces Faulty/Poor Decision Recommendation	Biased outcomes/predictions
	Skewed outcomes/predictions
	Uncertain outcomes/predictions
Human–Machine Operation Issues	Operators have lack of trust in the system
	Operators are overly trusting (overreliant) in the system
	Operators ignore the system
	Operators misunderstand the system recommendations/predictions
	Operators introduce errors into the system
System Under Attack (Cyber attack)	System is overtaken by adversary/adversary is controlling system
	System and its outcomes are corrupted by adversary
	Adversary jams or shuts down system
	Adversary gains access to system; decision information/knowledge is compromised

Table 6. Examples of Root Causes of AI System Failures. Source: Johnson (2021c).

Type of Root Cause	Root Cause Examples
Issues within the training datasets	Biased training datasets
	Incomplete training datasets
	Corruption in the training datasets
	Mis-labeled data
	Mis-associated data
	Lack of rare examples—data doesn't include unusual scenarios
	Unrepresentative datasets
Issues with the process of data validation	Poor data collection methods
	Poor data validation methods
	Improper data validation criteria
	Insufficient data validation
Issues with the ML algorithms	Underfitting in the model—when the model does not attain sufficiently low error on the training data
	Overfitting in the model—when the model presents very small error on the training data, but fails to generalize to new data
	Cost function algorithm errors—when the trained model is optimized to the wrong cost function
	Wrong algorithm—when the training data is fit to the wrong algorithmic approach or mathematical model
Issues with the operational datasets	Uncertainty/error in the operational datasets (Epistemic uncertainty)
	Corruption in the operational data
	Introduction of datatypes that the AI system is not designed to handle
	The pace of the situation overwhelms the human–machine decision process
Operational complexity	The decision space overwhelms the decision process (the number of options is too large or a viable option does not exist)
Operator trust issues	Lack of explainability
	Lack of confidence
	Overreliance
	Insufficient operator training or experience with system
Operator induced error	Inverse trust issues in which the AI system loses "trust" in the human operator or identifies operator problems
	Operator misuses the system accidentally or intentionally
	Operator fails their part in the decision process accidentally or due to being overwhelmed, negligent, or confused
Adversarial attacks	Hacking
	Deception
	Inserting false or corrupt data
	Gaining control of the AI system

Potential for failures of AI systems need to be identified early in the design process in order to mitigate catastrophic consequences from being a possibility.

D. SAFETY RISK ASSESSMENT

There are several organizations responsible for assessing the safety of weapon systems to include the various Program Offices and each of the military services for their particular programs. One such organization is Naval Ordnance Safety and Security Activity (NOSSA). As a field activity under NAVSEA, they are responsible for weapon system safety and software safety with regards to policy, procedure and design criteria (Naval Sea Systems Command 2021). They participate in many technical boards and panels but the one related to this project is the Software System Safety Technical Review Panel (SSSTRP). This technical review panel supports the Weapon System Explosives Safety Review Board (WSESRB). The objective of the SSSTRP is to provide a thorough review of the software control of weapon systems (Shampine 2010). The panel reviews the Technical Data Package provided for WSESRB review.

The team needed to reference a couple process documents to complete the analysis of the three operational scenarios in this project. First, the *Joint Services Software Safety Authorities (JS-SSA) Software System Safety Implementation Process and Tasks Supporting MIL-STD-882E (Implementation Guide)* (2016) provided tasks and subtasks at the system level to ensure the proper level of rigor is used to design safe software and to define needed safety requirements supporting all software builds from design through test validation phases. Next, *NIST Special Publication 800-37 Revision 2* (2018) outlines the Risk Management Framework (RMF) and provides guidance in which to apply the RMF process to information systems and organizations.

Both documents include many tasks and subtasks that should be completed to provide for development of a safe system. The team identified tasks that were within the scope of this study to complete the analysis for each of three scenarios while other tasks were outside the scope. The first task for each scenario was to perform an in-depth hazard analysis which identified hazard failure modes and causes for the system. Documentation of the functional failure consequences would follow. Next, a safety risk assessment was

performed on the documented hazards. Mitigations were suggested for each of the hazards that could be performed to ensure a safe system.

The National Institute of Standards and Technology (NIST) Information Technology Laboratory (ITL) is a technical leader for the nation's measurement and standards infrastructure (NIST 2018). By developing such resources as tests (along with the methods by which they are performed), proofs of concept, references, and technical analysis, the NIST ITL manages the standards and guidelines for security engineering within federal information systems (2018). NIST's publications provide references and guidelines to abide by for security and privacy for industry government, and academic information systems. NIST, in its partnership with the Department of Defense, the Office of the Director of National Intelligence, and the Committee on National Security Systems, developed the Risk Management Framework to refine security for DOD information systems, enhance the risk management processes, and create standardization and consistency among organizations (NIST 2018).

The RMF prepares organizations to implement and execute the framework activities for authorization through the use of continuous tracking practices, which enables decision-makers to manage the risk efficiently and effectively, and implement privacy and security through the system development life cycle (NIST 2018). RMF execution links risk management process at system and organizational levels, as well as establishes accountability and responsibilities for the controls to be implemented within information systems. The RMF process ensures the software is safe from a security perspective. The weapon system software must be evaluated under this RMF process in order to receive an authority to operate (ATO).

Figure 16 depicts the RMF process steps. The process begins with the Prepare step, but the remaining steps can be completed in any order. Some steps are performed at the organization level while others are executed at the system level. Each step has several associated tasks. See the example in Table 7.

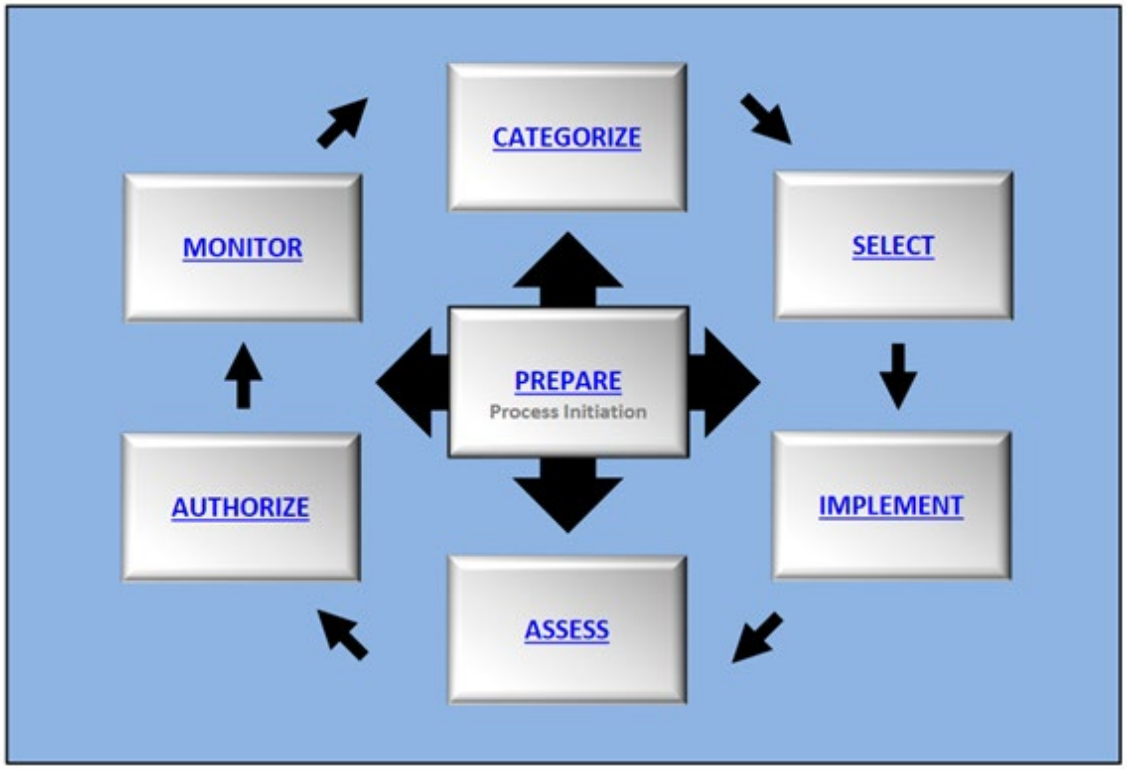

Figure 16. RMF Process Steps. Source: NIST (2018).

Table 7. RMF Tasks. Source: NIST (2018).

RMF TASKS	PRIMARY RESPONSIBILITY	SUPPORTING ROLES
TASK R-1 **Authorization Package** Assemble the authorization package and submit the package to the authorizing official for an authorization decision.	• System Owner • Common Control Provider	• System Security Officer • System Privacy Officer • Senior Agency Information Security Officer • Senior Agency Official for Privacy • Control Assessor
TASK R-2 **Risk Analysis and Determination** Analyze and determine the risk from the operation or use of the system or the provision of common controls.	• Authorizing Official or Authorizing Official Designated Representative	• Senior Accountable Official for Risk Management or Risk Executive (Function) • Senior Agency Information Security Officer • Senior Agency Official for Privacy
TASK R-3 **Risk Response** Identify and implement a preferred course of action in response to the risk determined.	• Authorizing Official or Authorizing Official Designated Representative	• Senior Accountable Official for Risk Management or Risk Executive (Function) • Senior Agency Information Security Officer • Senior Agency Official for Privacy • System Owner or Common Control Provider • Information Owner or Steward • Systems Security Engineer • Privacy Engineer • System Security Officer • System Privacy Officer
TASK R-4 **Authorization Decision** Determine if the risk from the operation or use of the information system or the provision or use of common controls is acceptable.	• Authorizing Official	• Senior Accountable Official for Risk Management or Risk Executive (Function) • Chief Information Officer • Senior Agency Information Security Officer • Senior Agency Official for Privacy • Authorizing Official Designated Representative
TASK R-5 **Authorization Reporting** Report the authorization decision and any deficiencies in controls that represent significant security or privacy risk.	• Authorizing Official or Authorizing Official Designated Representative	• System Owner or Common Control Provider • Information Owner or Steward • System Security Officer • System Privacy Officer • Senior Agency Information Security Officer • Senior Agency Official for Privacy

The table of tasks also depicts who has primary responsibility and supporting roles in completing the task. This can be a lengthy process to complete, so the entire process is

not part of the scope of this project. Chapter IV contains a system level risk assessment of the failure modes and hazard analysis done in Chapter III.

III. SCENARIOS, FAILURE MODES, AND HAZARD ANALYSIS

This chapter tackles the first objective of identifying the safety risks related to the deployment of AI systems that support future automated tactical decision and mission planning aids. First, the team analyzes hazards for a generic AAMD engagement, utilizing the functional hierarchy. This analysis leads to three distinct use case scenarios for artificial intelligence / machine learning use in battle management aids, which the team evaluates for likely failure modes associated with each. The first use case scenario investigates trust deficit in an event in which ballistic missile defense (BMD) assets are on alert supporting a homeland defense mission. The second use case investigates the potential perils of training data provided to AI/ML on board a naval vessel in a ship self-defense scenario. The third use case explores an area defense scenario and the issues associated with the strategic vs. theater bias.

The generic AAMD engagement offers a look into common safety concerns for computer systems and the impact to AI/ML involvement. Currently, AAMD takes place without AI/ML. Artificial intelligence contributes to this mission to better integrate capabilities, such as tasking, prioritizing threats, scalable pre-planned responses, and gathering data for calculating or learning. Tasking refers to the optimal, coordinated use of each system within the AAMD context. In a semi-automatic setting, AI advises and assists the warfighter (WF). Semi-autonomy can range from purely advising the WF to conditionally assisting/executing operations (i.e., time-sensitive situations). In an automatic setting, AI executes operations on its own. A fully automatic setting differs entirely from current CONOPS and introduces unique safety concerns. For the purposes of this analysis, AI involvement will be primarily focused on a semi-automatic setting, such as protocols between user and AI.

The generic AAMD engagement acts as a baseline for the analysis of each AI/ML BMA scenario. A context diagram provides a connection between each scenario and the functions discussed in the generic AAMD engagement. The hazard analysis of the three scenarios utilizes a fault tree to identify possible failure modes. The deductive procedure of a fault tree analysis determines hardware failures, software failures, and human errors

that could cause specific mishaps within the context of the scenario, while providing visualization of the failure modes, hazards, and mishaps.

A. GENERIC AAMD ENGAGEMENT

The team used Innoslate, a model-based systems engineering (MBSE) tool, to model the generic AAMD engagement functions. This aids in identifying general safety concerns. Shown in Figure 17, a functional hierarchy narrowed down the engagement to four main functions: sense, communicate, engage, and kill assessment. Figure 17 is an initial representation of a simple AAMD engagement. The activity diagram, in Figure 18, represents how these four main functions interact with each other, along with general inputs and outputs.

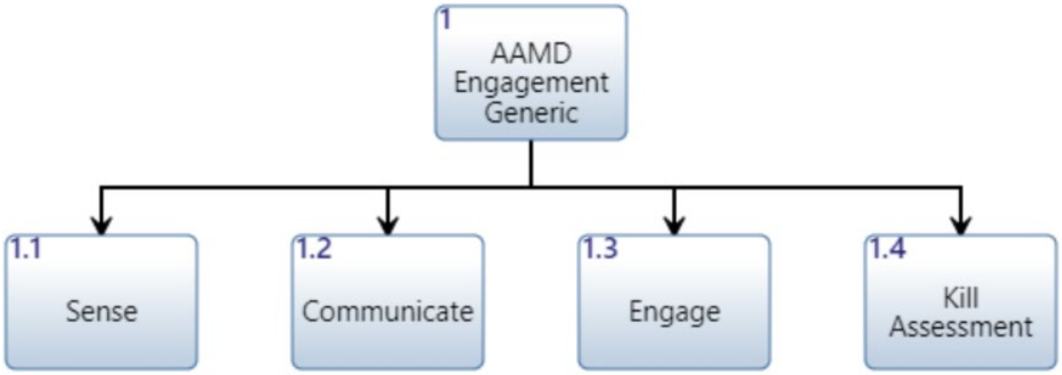

Figure 17. Generic AAMD Engagement Functional Hierarchy (Level 1)

The sense (1.1) function contains the detect, track, and identify sub functions for modeling simple radar functionality. Communicate (1.2) contains the functions necessary to support decision making. Under communicate, the output is an engagement command that feeds the Engage (1.3) function. The Engagement function models the simple subfunctions of selecting a shooter within the architecture, sending a launch command, then launching an interceptor (generic). Finally, the kill assessment (1.4) verifies the threat status for refire or engagement completion. In Innoslate, the kill assessment contains the "roll of the dice" based on a pre-determined probability of kill. If a kill is achieved, a kill

message goes out to the architecture. If an interceptor hit is not achieved, the model will continue shooting interceptors until the threat is killed.

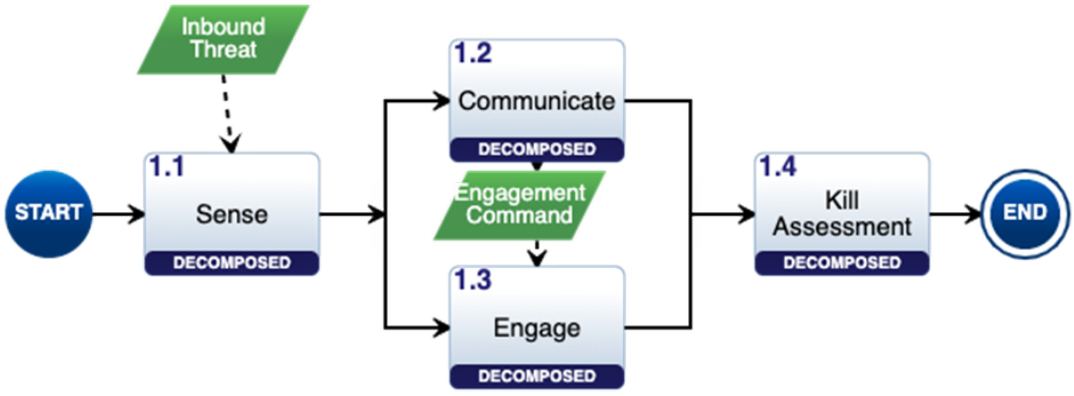

Figure 18. Generic AAMD Engagement Action Diagram

Details of the four functions aid in identifying safety concerns. The associated subfunctions and systems facilitate characterizing the failures associated with each safety concern. The types, as defined in Table 8, categorize the failures by the potential causes for it to occur.

Table 8. AAMD Engagement Failure Category Types

Failure Type	Definition	Examples
Operational	Failure of system operation or system to system operation	Internal sensor function failure, launcher malfunction
AI/ML Programming	Incorrect/unintended error in AI/ML programming	Identify hostile threat as non-hostile, unable to process multiple threats
Adversarial Attack	Direct attack or manipulation by adversary	C2 network hacking, insider threat, enemy causes ML recognition mistake
Human-Machine Interaction (HMI)	Errors with user interaction with the system(s) (AI interaction focused)	Interface issues, interpretation error, lack of trust in AI/ML

1. AAMD Engagement – Sense

The sense function decomposes, as seen in Figure 19, to three main subfunctions; detect, track, and identify. The activity diagram, in Figure 20, implies that these functions are done in series. Sensors detect an inbound threat, collect tracking information, and collect data to identify the incoming threat. Track and identify functions may be performed by several subsystems, in coordination with AI/ML. The AI/ML BMA processes the collected sensing data to actively track and identify threats. The WF monitors subsystems performing functions 1.1.1, 1.1.2, and 1.1.3. The threat indicates adversarial involvement.

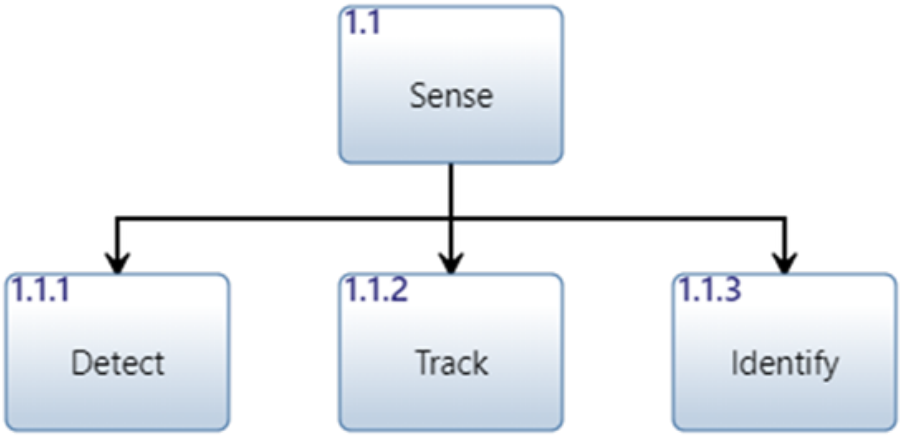

Figure 19. AAMD Engagement Hierarchy Diagram – Sense

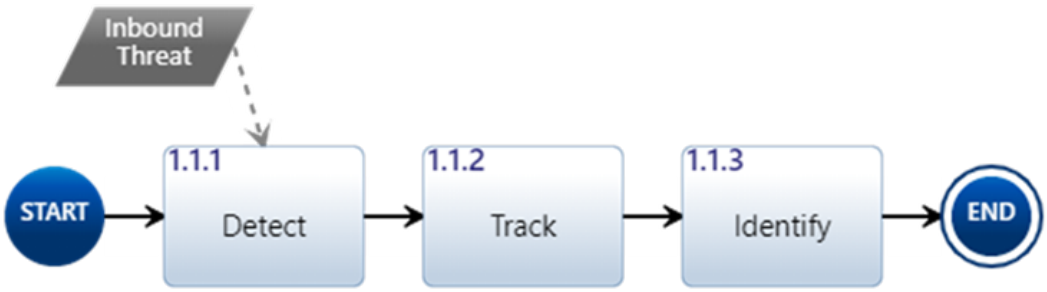

Figure 20. AAMD Engagement Activity Diagram – Sense

Table 9 names seven safety concerns with their associated functions, systems, and failure types. Although Figure 20 only shows one inbound threat, the safety concerns consider the system of systems handling multiple threats. The AI/ML BMAs have the capacity to handle that information more effectively than the WF alone. Artificial intelligence also manages the tracking and identification of functions. However, that introduces AI/ML programming as a main failure type for the sense function. The WF mainly serves as a supervisor for the HMI failures.

Adversarial attacks associated with S-6 and S-9 include direct and indirect attacks on systems. A direct attack physically takes out a sensor or AI/ML BMA communication line. Indirect attacks on AI/ML BMA involve attacks on the initial programming or ML. For example, an adversary sends a small swarm of non-hostile drones, in an effort to train ML to recognize this as a non-hostile. After some time, an adversary sends a similar swarm of hostiles that AI/ML BMA mistakenly marks as non-hostile, giving the hostiles a better probability of success.

Table 9. AAMD Engagement Safety Concerns – Sense

ID	Safety Concern	Related Function(s)	Related System(s)	Failure Type(s)
S-1	Failure to detect threat	1.1.1	Sensor, AI/ ML BMA	Operational, AI/ML Programming
S-2	Failure to detect multiple threats	1.1.1	Sensor, AI/ ML BMA	Operational, AI/ML Programming
S-3	Conflicting/confusing detections for multiple threats	1.1.1, 1.1.2	Sensor, AI/ ML BMA, WF	Operational, AI/ML Programming, HMI
S-4	Conflicting/confusing detection and/or tracking	1.1.1, 1.1.2	Sensor, AI/ ML BMA, WF	Operational, AI/ML Programming, HMI
S-5	Lack of data for further calculations (i.e., trajectory)	1.1.2	Sensor, AI/ ML BMA	Operational, AI/ML Programming
S-6	Loss of threat tracking	1.1.2	Sensor, AI/ ML BMA	Operational, Adversarial Attack
S-7	Conflicting/confusing tracking data	1.1.2	Sensor, AI/ ML BMA	Operational, AI/ML Programming
S-8	Misidentify threat type (weapon type)	1.1.3	Sensor, AI/ ML BMA, WF	Operational, AI/ML Programming, HMI
S-9	Misidentify friendly for threat	1.1.3	AI/ML BMA, WF, Adversary	AI/ML Programming, HMI, Adversarial Attack

2. AAMD Engagement – Communicate

The hierarchy diagram in Figure 21 contains subfunctions between AI and WF to develop a COA. As seen in Figure 22, function 1.2 permits a selection between an engagement under full automation or with the WF in the loop. Full automation simply allows for faster processing of the engagement while the selection of WF, in the loop, slows the process down – given that the WF follows a CONOPs of certain pre-planned responses. Alternatively, AI carries out pre-planned responses more quickly, but scalability comes down to the programming data. Function 1.2 generates and outputs the COA input for function 1.3.

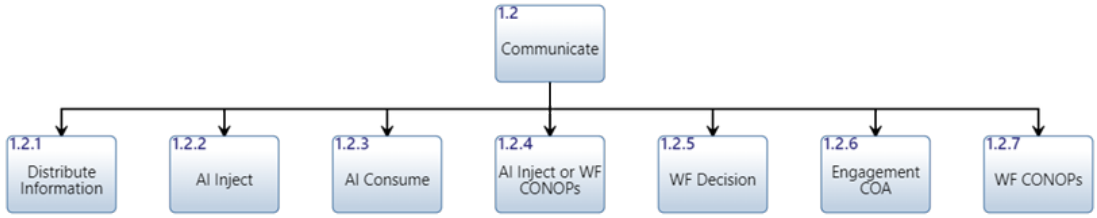

Figure 21. AAMD Engagement Hierarchy Diagram – Communicate

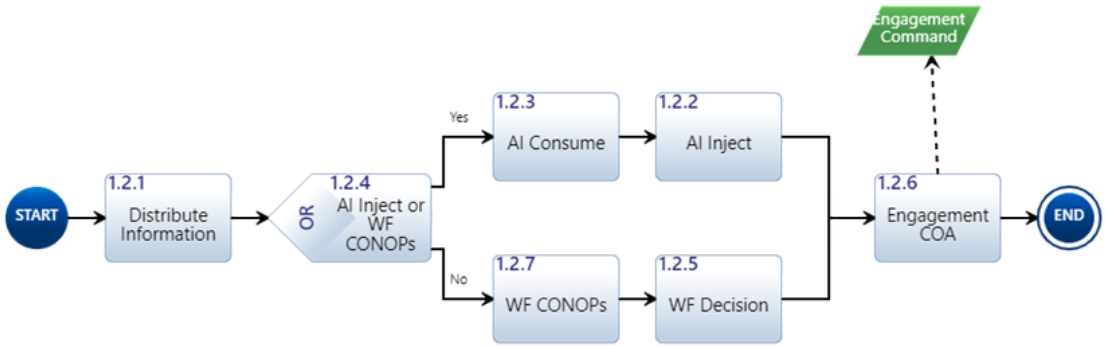

Figure 22. AAMD Engagement Activity Diagram – Communicate

This function relies on the WF CONOPS and AI/ML programming to develop a COA. As noted in Table 10, the WF and AI/ML BMA relate to the majority of the safety concerns. Thus, HMI and AI/ML programming become the main failure types for this function's safety concerns. The safety concerns also involve the cooperation of the WF and AI/ML BMA.

Table 10. AAMD Engagement Safety Concerns – Communicate

ID	Safety Concern	Related Function(s)	Related System(s)	Failure Type(s)
C-1	Unable to distribute information	1.2.1	C2 network, system to system interface	Operational
C-2	Miscalculate threat impact	1.2.2, 1.2.3	AI/ML BMA	AI/ML Programming
C-3	Delay in AI recommendation (time-sensitive situation)	1.2.1, 1.2.2, 1.2.3	C2 network, AI/ML BMA	Operational, AI/ML Programming
C-4	Ineffective/inefficient prioritization of threats (to minimize impact)	1.2.2, 1.2.5, 1.2.6, 1.2.7	AI/ML BMA, WF	AI/ML Programming, HMI
C-5	Insufficient/outdated CONOPS	1.2.7	WF	Operational
C-6	Outdated COA chosen (time-sensitive COA recommendation)	1.2.2, 1.2.4, 1.2.6	AI/ML BMA, WF	AI/ML Programming, HMI
C-7	Mistrust of AI recommendation	1.2.2, 1.2.4, 1.2.5	AI/ML BMA, WF	AI/ML Programming, HMI

3. AAMD Engagement – Engage

In Figure 23, function 1.3 decomposes simply to three subfunctions, engaging the targeted threat in accordance with the engagement output from function 1.2. The subfunctions follow each other in series to carry out the COA input. Within the AAMD systems of systems, systems communicate with each other throughout the activity diagram in Figure 24. The level of AI autonomy (AI and WF involvement) varies for function 1.3, as intended for this analysis.

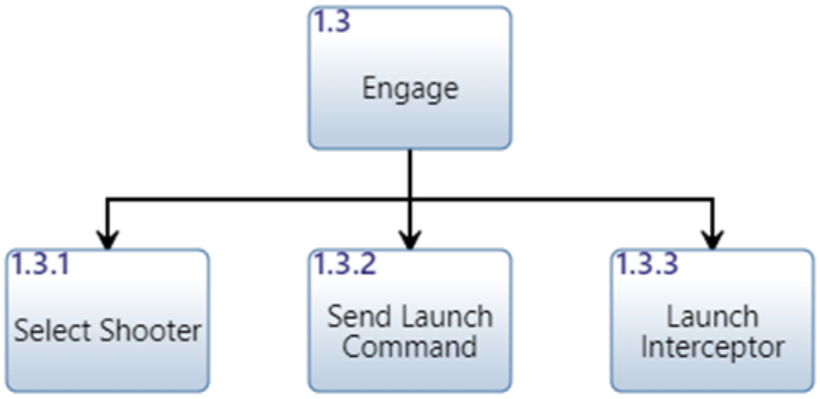

Figure 23. AAMD Engagement Hierarchy Diagram – Engage

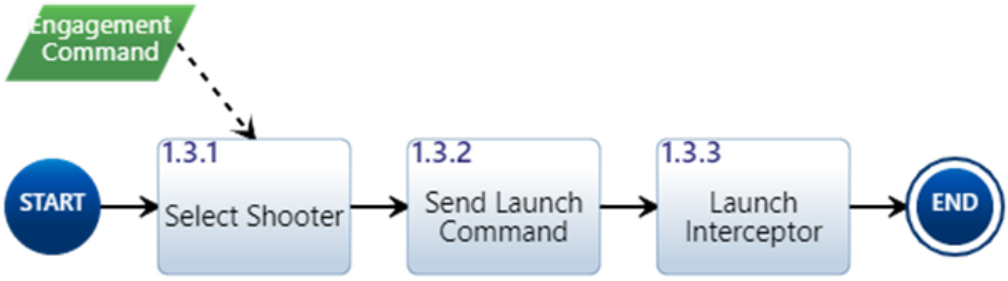

Figure 24. AAMD Engagement Activity Diagram – Engage

The WF or AI executes engagement subfunctions, driving the associated failure types (HMI and AI/ML programming) for many safety concerns in Table 11. Failures at this point result in direct hazards due to interceptor involvement. Therefore, the safety concerns point to the need for a manual override function as an option, for future models.

Table 11. AAMD Engagement Safety Concerns – Engage

ID	Safety Concern	Related Function(s)	Related System(s)	Failure Type(s)
E-1	Selection not ready for launch or offline	1.3.1, 1.3.2	C2 Network, Weapon System, AI/ ML BMA	Operational, AI/ML Programming
E-2	Prolonged interceptor offline to online protocol	1.3.1, 1.3.2, 1.3.3	Weapon System	Operational
E-3	Engagement command error	1.3.1 (input)	C2 Network, WF, Adversary	Operational, Adversarial Attack, HMI
E-4	Physical launching error	1.3.3	Weapon System	Operational
E-5	Unidentifiable launching error	1.3.3	Weapon System, AI/ ML BMA, WF	Operational, AI/ML Programming, HMI
E-6	Incorrect interceptor launched (different from command)	1.3.1, 1.3.2, 1.3.3	Weapon System, AI/ ML BMA, WF, Adversary	Operational, Adversarial Attack, HMI
E-7	Not following command recommendation (possible insider threat)	1.3.1, 1.3.3	WF, Adversary	Adversarial Attack

4. AAMD Engagement – Kill Assessment

Function 1.4 follows up on the status of function 1.3. The weapon launched in function 1.4 needs to be tracked for threat status. A kill or miss message, in Figure 25, ends the initial engagement. However, a miss message triggers another engagement sequence. This loop is not shown in Figure 26 since refire calls for a loop of the entire engagement. The decomposition in Figure 25 supports attributing specific function safety concerns and related refire protocol to functions 1.4.2, 1.4.3, and 1.4.4.

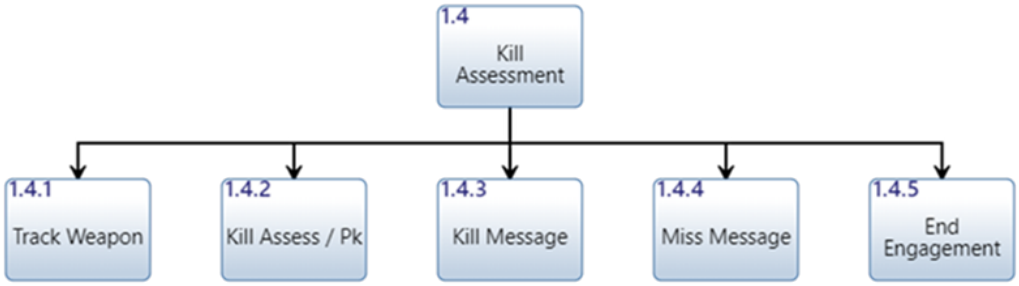

Figure 25. AAMD Engagement Hierarchy Diagram – Kill Assessment

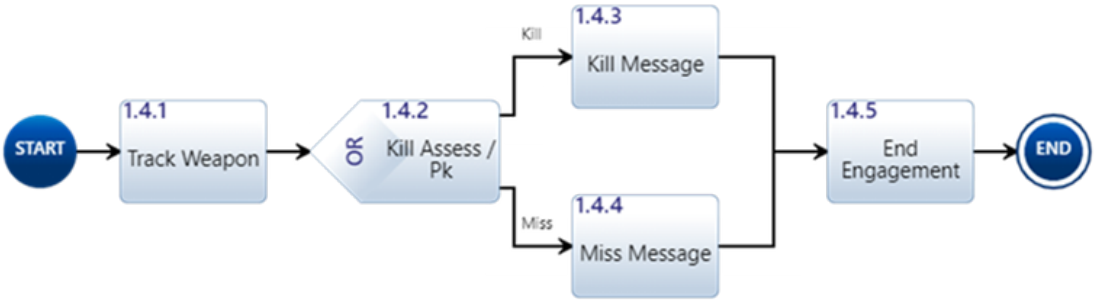

Figure 26. AAMD Engagement Activity Diagram – Kill Assessment

Function 1.4 solely relies on system operation and AI/ML programming. Artificial intelligence consumes the information from other systems to determine a kill. If the other systems operate normally, AI manages the kill assessment determination. Safety concerns K-3 and K-7, in Table 12, relate to HMI with the WF. This is due to message clarity (functions 1.4.3 and 1.4.4) and/or interface design.

Table 12. AAMD Engagement Safety Concerns – Kill Assessment

ID	Safety Concern	Related Function(s)	Related System(s)	Failure Type(s)
K-1	Failure to track weapon(s) (single/ multiple threats)	1.4.1	Sensor, AI/ ML Programming	Operational, AI/ML Programming
K-2	Failure to assess kill	1.4.2	Sensor, AI/ ML Programming	Operational, AI/ML Programming
K-3	Confusing/conflicting messages (multiple threats)	1.4.2, 1.4.3, 1.4.4	AI/ML BMA, WF	AI/ML Programming, HMI
K-4	Misidentified miss as kill	1.4.2	Sensor, AI/ ML BMA	Operational, AI/ML Programming
K-5	Delayed miss message (time sensitive situation for refire)	1.4.2, 1.4.4	C2 Network, AI/ML BMA	Operational, AI/ML Programming
K-6	Kill assessment loop takes too long for effective refire	1.4.2, 1.4.4	Sensor, C2 Network, AI/ ML BMA	Operational, AI/ML Programming
K-7	Unknown/confusing protocol for refire	1.4.2, 1.4.4, 1.4.5	AI/ML BMA, WF	AI/ML Programming, HMI

5. Common AI System Hazards

Computer systems produce their own hazards without AI involvement. They rely on power sources, data consumption, user competency, and network communications. Table 13 reveals common computer system hazards and their impact if the system was AI/ ML.

Table 13. Common Computer AI System Hazards

ID	Hazard	AI/ML Impact
H-1	Natural Disaster	Systems unusable/destroyed; AI has limited/no information to consume and limited assets to negate threat, effective COA from AI decreases
H-2	Power Loss	Systems offline or in manual mode and unable to communicate directly; AI/ML unusable
H-3	Network Related Adversarial Attack	C2 network communication affected; AI/ML becomes untrustworthy
H-4	System Component Failure	AI programming needs to be aware of system component failure and calculate COA accordingly (may be limited or no assets to use)
H-5	Corrupt/Incorrect Data	AI/ML becomes untrustworthy
H-6	User Error/Lack of Knowledge or Training	AI/ML contribution increases in value (threat recognition/calculations, recommendation on COA, autonomy, etc.)
H-7	Out of Date System	AI/ML outdated for optimal use; AI produces less effective/useful recommendations
H-8	Insider Threat	All systems at risk to be compromised; AI/ML becomes untrustworthy
H-9	Weak Access Controls	Security of systems at risk; escalated risk for corrupt system or adversarial attacks on AI/ML
H-10	Encryption Failure	Security of systems and COA at risk; escalated risk for adversarial attacks on AI/ML

In Table 14, the team captured four key AI/ML problems from this generic AAMD engagement analysis. The team developed three scenarios to examine A-1, A-2, A-3 further for explicit failure modes. A-4 notes a generic concern which applies to all scenarios. Therefore, it is integrated into the failure mode analysis of each scenario instead of given its own scenario.

Table 14. AAMD Scenario Hazards

ID	Scenario Concern	Failure Type(s)	Failure Modes
A-1	AI/ML BMA recommendation on engagement differs from WF CONOPS and Training, Tactics, and Procedures	AI/ML Programming, HMI	See BMD Scenario (scenario 1)
A-2	Threat has been mis-identified; Blue Forces are unaware of threat; have wrong information about threat; or confusing/conflicting information about threat	AI/ML Programming, Adversarial Attack	See Ship Self Defense Scenario (scenario 2)
A-3	AI/ML BMAs at different Command and Control nodes make different recommendations	Operational, AI/ML Programming, HMI	See Area Defense Scenario (scenario 3)
A-4	AI/ML BMA's recommended course of action is wrong/ineffective/inefficient	AI/ML Programming	Programming and data inputs for AI/ ML; training data

B. BALLISTIC MISSILE DEFENSE

1. Scenario Description

This first scenario involves the strategic mission of defending the United States from ballistic missile threats (particularly those armed with weapons of mass destruction such as nuclear payloads). Ballistic missile defense assets are on alert supporting a homeland defense mission. A strategic, intercontinental ballistic missile (ICBM) threat is headed towards the continental U.S. with a suspected nuclear payload.

As the threat missile makes its way towards its target, BMD assets sense the threat and begin the tracking process. These assets communicate the threat through the appropriate command and control ballistic missile communications (C2BMC) channels. Appropriate leadership is notified and through established standard operating procedures (SOPs), they devise a course of action (COA) to combat the threat. Simultaneously, AI assets within the C2BMC channels consume the information, devise a separate (though not

necessarily different) COA, and inject it into the solution. The WF orient on the COA and engage the threat by way of either the AI's recommendation, or their own concept of operations (CONOPS). The context diagram in Figure 27 depicts an example of this engagement sequence. This scenario focuses on WF CONOPs and Strategic AI boxes and their influences on the GMD BMA box.

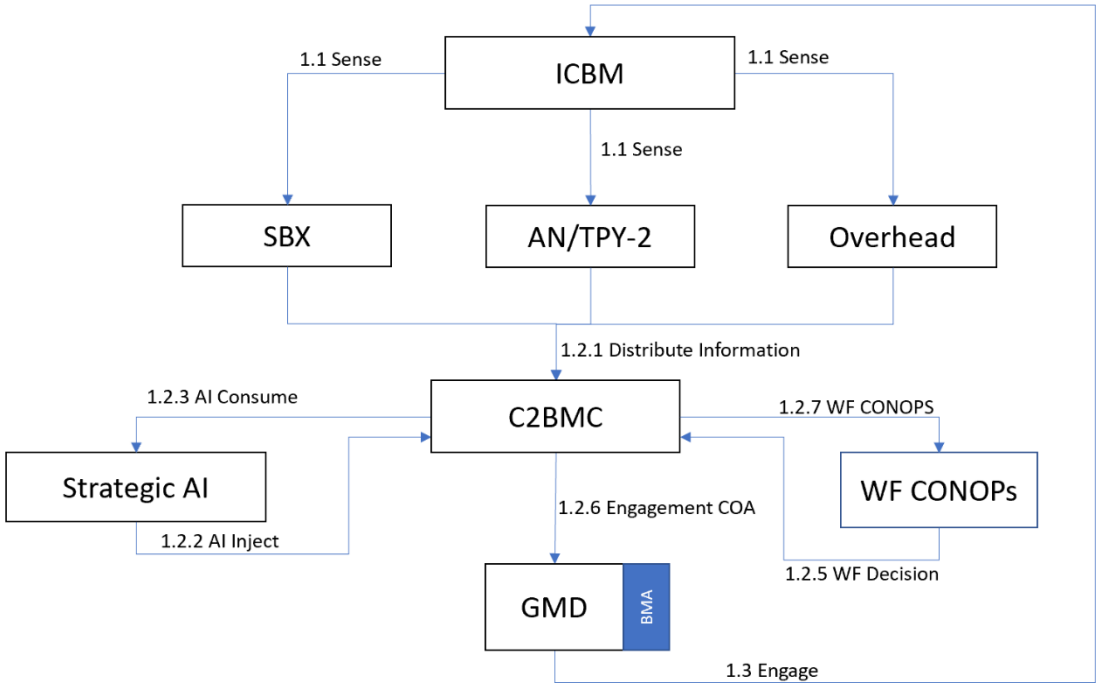

Figure 27. BMD Context Diagram

The context diagram offers an initial look at potential failure modes. This specific scenario is a broad look at AI/ML used as a BMA. There are many problems that could lead to a failure.

2. Failure Modes and Hazard Analysis

There are too many potential threats and failure modes to address with regards to a BMD scenario in this paper. The team focused the scenario analysis on the issue of differing engagement recommendations between AI/ML BMA and WF CONOPS/TTPs. Per *MIL-STD-882, Rev E*, failure modes are identified by tracing the primary failure paths

leading to a hazard, mishap, and mishap effect (DoD 2012). The main mishap effect concerns loss of life and assets due to the impact of the incoming threat (nuclear payload) and possible collateral damage from the countermeasures launched to neutralize the threat. The hierarchy diagram in Figure 28 depicts the potential hazards and associated failure modes, which are represented by the 15 red boxes.

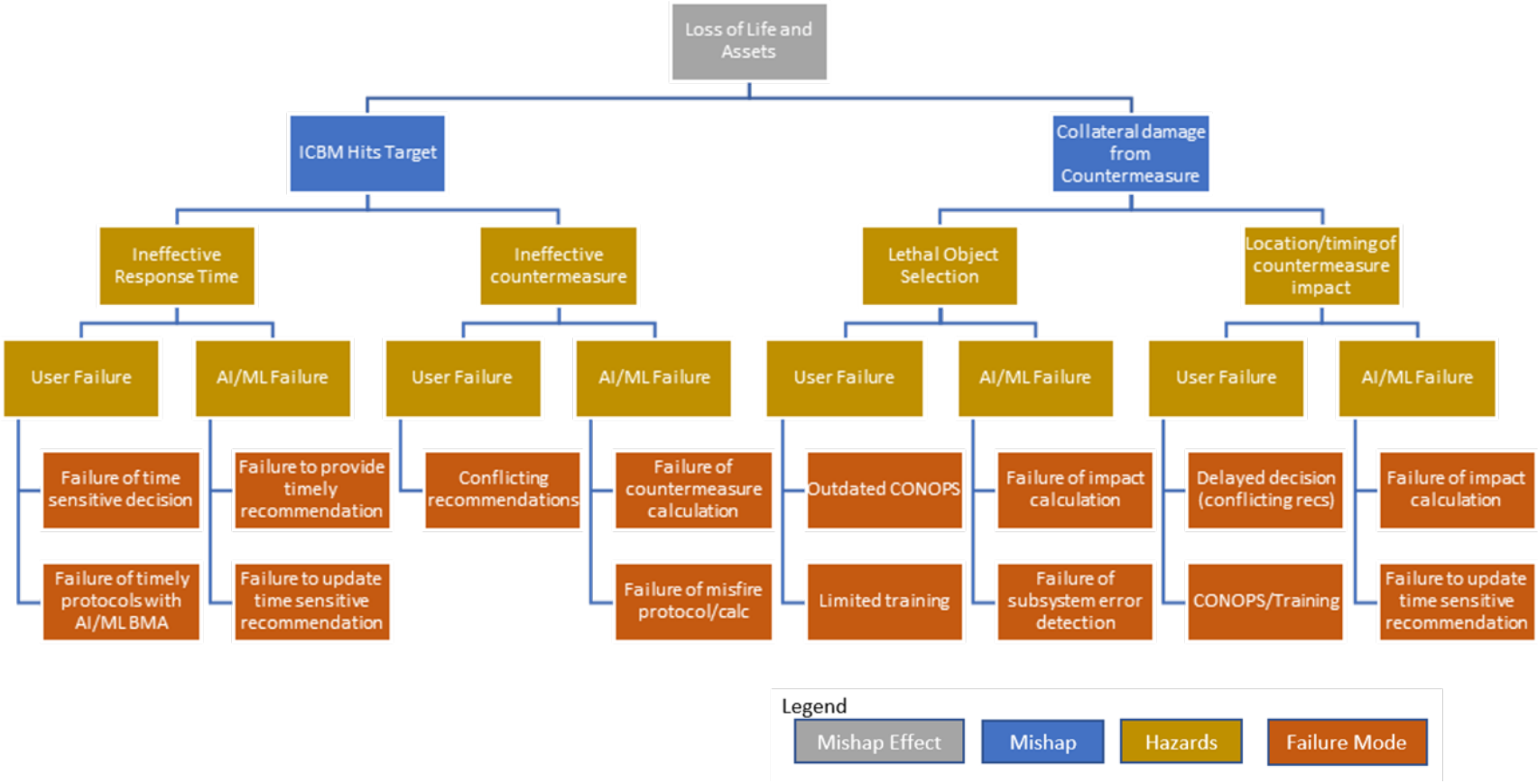

Figure 28. BMD Hazard Failure Mode Tree (WF Trust Deficit)

Again, the focus for this scenario is the WF's trust in AI/ML BMA and its recommendations on engagement. Therefore, the team needs to compare the failures caused by the user (WF) against the failures caused by AI/ML, to analyze the safety risks. As shown in Figure 28, the failure modes are categorized in this manner, for each identified hazard. The categorized types of these failure modes vary between AI/ML programming, operational, and HMI. Table 15 concludes that this scenario relies on effective cooperation between the systems, including the WF.

Table 15. BMD (WF Trust Deficit) Failure Mode Summary

Failure Type(s)	**Failure Mode**	**Related Entities**	**Related Functions**
AI/ML Programming	AI failure of countermeasure calculation	Strategic AI, GMD BMA	1.2.2, 1.2.3
AI/ML Programming	AI failure of impact calculation	Strategic AI, GMD BMA	1.2.2, 1.2.3, 1.2.6, 1.3
HMI	WF failure of time sensitive decision	WF CONOPs	1.2.5
Operational	Outdated CONOPs	WF CONOPs	1.2.5, 1.2.7
Operational	Incorrect CONOPs/Training	WF CONOPs	1.2.5, 1.2.7
Operational	AI failure to provide timely recommendation	Strategic AI, C2BMC	1.2.2, 1.2.3
Operational	AI failure of subsystem error detection	Strategic AI, C2BMC, SBX, AN/ TYP-2, Overhead, GMD BMA	1.2.1, 1.2.3
Operational, AI/ML Programming	Conflicting recommendations	WF CONOPs, Strategic AI	1.2.2, 1.2.5
Operational, AI/ML Programming	AI failure to update time sensitive recommendation	Strategic AI, C2BMC, WF CONOPs	1.2.2, 1.2.3
Operational, AI/ML Programming	AI failure of misfire protocol/ calculation	Strategic AI, SBX, AN/ TPY-2, Overhead	1.1, 1.2.1, 1.2.2, 1.2.3, 1.2.6, 1.3

Failure Type(s)	Failure Mode	Related Entities	Related Functions
Operational, AI/ML Programming, HMI	WF delayed decision (conflicting recommendations)	WF CONOPs, C2BMC, GMD BMA	1.2.2, 1.2.5
Operational, HMI	WF failure of timely protocols with AI/ML BMA	WF CONOPs, Strategic AI, C2BMC, GMD BMA	1.2.2, 1.2.3, 1.2.5, 1.2.6, 1.2.7
Operational, HMI	Limited training	WF CONOPs	1.2.5

C. SHIP SELF DEFENSE

1. Scenario Description

This scenario describes the defense of a naval warship. During the software development phase, machine learning algorithms were provided a host of threats in order to build a Combatant Command-agnostic AI/ML BMA. This training data ranged over various tactical-level threats. An Aegis BMD destroyer is patrolling contested waters. Artificial intelligence/machine learning has observed swarms of unmanned aerial vehicles (UAV) and has a model of what the threat looks like. All previous encounters have been non-hostile and suspected of being surveillance drones watching the ship. A much smaller swarm of UAV approaches the ship. Based on previous data, AI/ML does not recognize the swarm as a threat, and the ship is attacked.

Figure 29 follows the generic functions for an engagement against a swarm of UAVs by the Aegis ship. Local and organic sensors will sense the threat complex and help identify the threat. Tracking and targeting data is communicated through the command and control systems. Given the limited training data or observed data by the AI/ML, the system does not recognize the threat swarm and does not provide an appropriate recommendation.

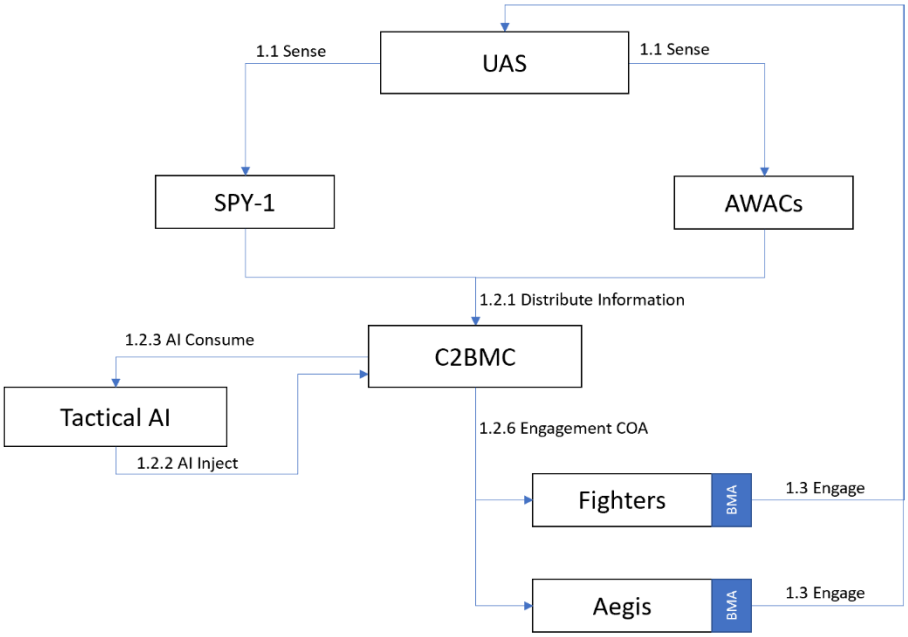

Figure 29. Ship Self Defense Context Diagram

2. Failure Modes and Hazard Analysis

This scenario offers two main mishaps to discuss regarding training data: hostile UAVs successfully attack Aegis or Aegis launches an attack on non-hostiles. Both are rooted in the training data that was used to program the AI BMA and the ongoing machine learning programming. In a successful hostile attack, as described in the scenario, AI/ML BMA misidentifies the small swarm of UAVs as a non-threat and does not recommend engagement. Figure 30 identifies the failure modes for the misidentification of the hostile UAV swarm. This mishap may also be caused by an ineffective response recommended by AI/ML BMA, which correctly identifies the UAV swarm. This hazard is depicted in Figure 31 as well, for completeness.

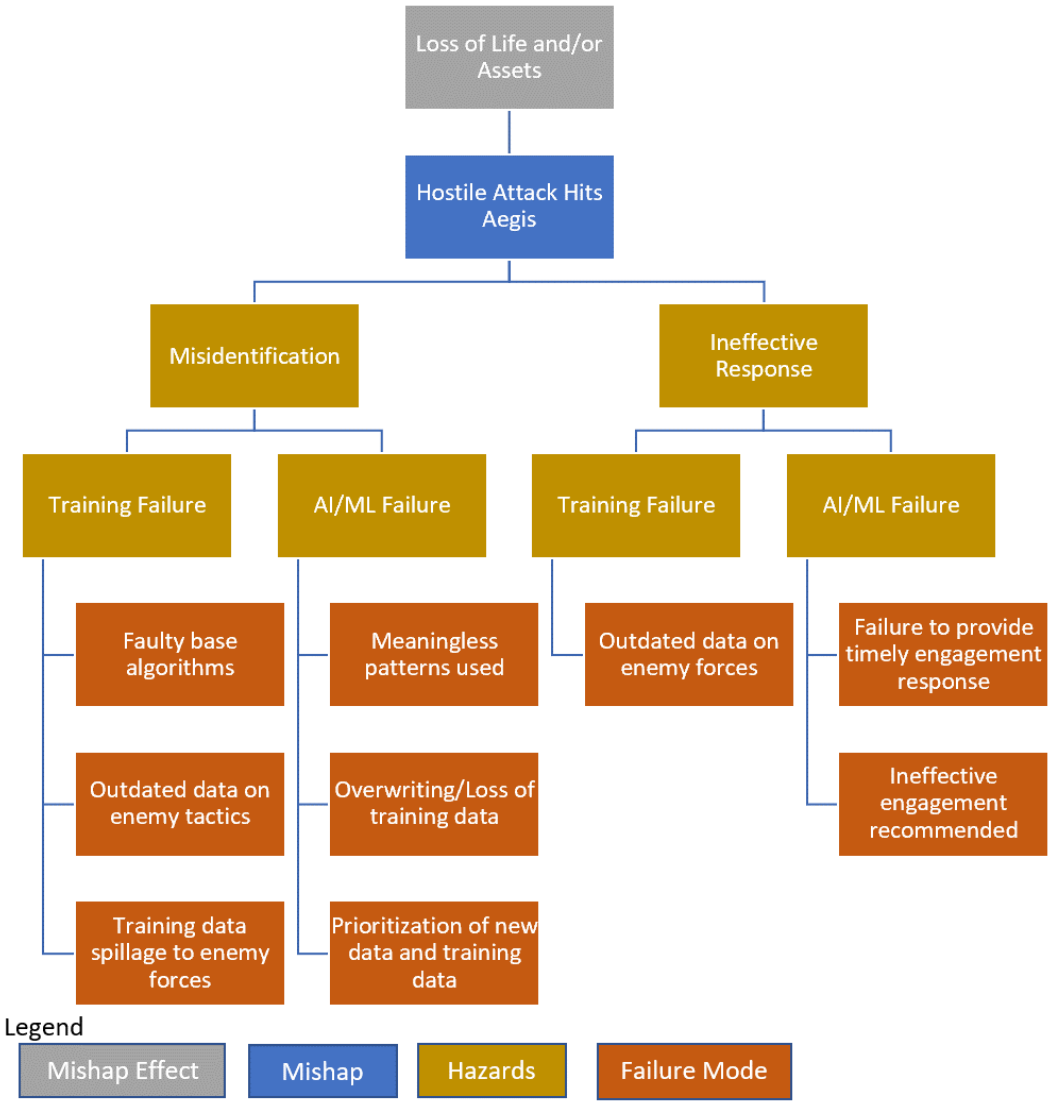

Figure 30. Hazard Failure Mode Tree for Incoming Hostile Attack (Training Data)

On the other hand, misidentification could be the opposite; AI/ML BMA misidentifies a non-hostile swarm of UAVs as a threat and launches an attack. This may not pose an inherent safety concern. However, the launched attack on non-hostiles may cause follow-on enemy attacks where loss of life and/or assets becomes a higher probability. Figure 31 explores this concept of attacking non-hostiles and associated failure modes. Failure modes are categorized as training failure or AI/ML failure.

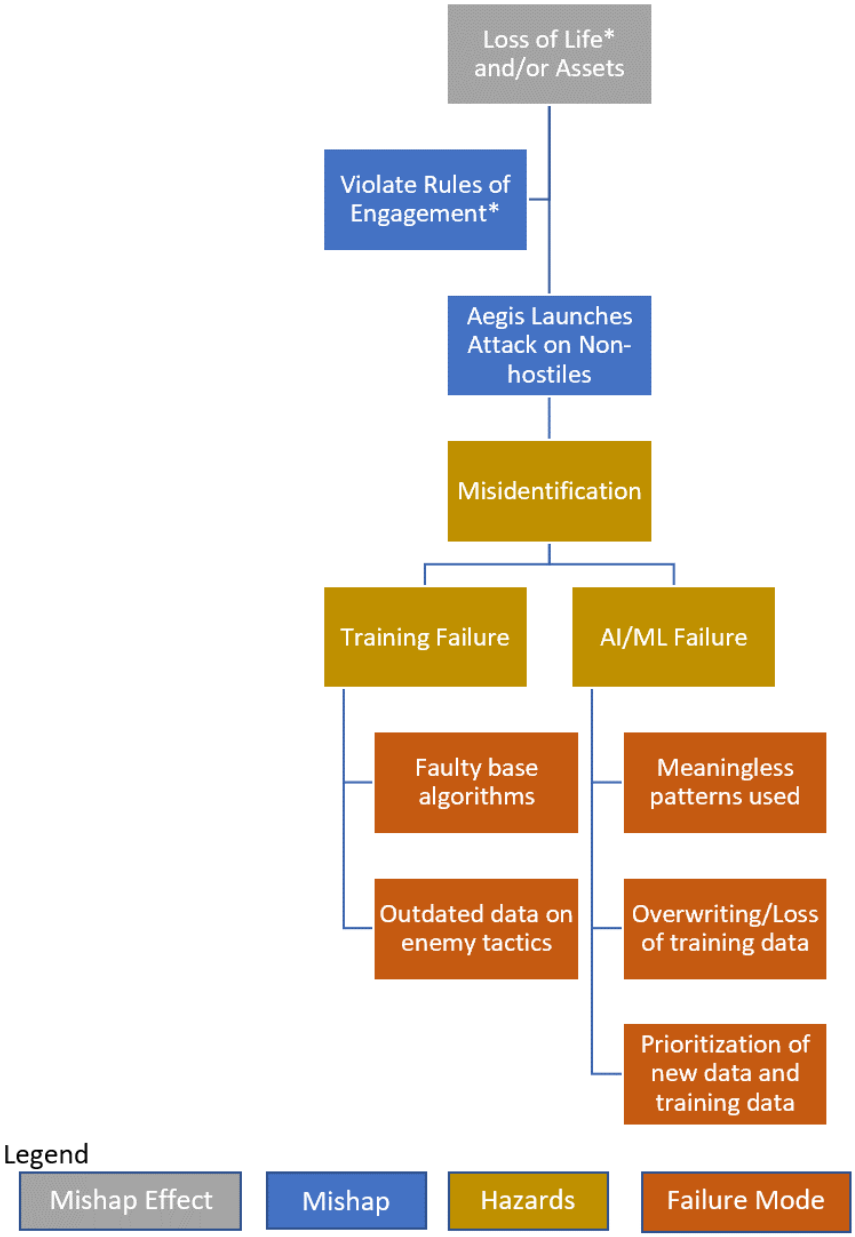

Figure 31. Hazard Failure Mode Tree for Attack on Non-hostiles (Training Data)

This scenario focuses on training data and machine learning. Almost every identified failure mode involves AI/ML programming, as indicated in Table 16. Operational failure modes encompass errors in how AI/ML is intended to work within this system of systems. Adversarial attacks, in this scenario, include direct attacks to the AI entities but also passive attacks or manipulation of AI functions. For example, an adversary

purposefully sends small swarms of surveillance UAS to train the AI/ML to identify as non-hostile. The adversary now sends a hostile UAS swarm of comparable size/ characteristics that AI identifies as non-hostile.

Table 16. Ship Self Defense (Training Data) Failure Mode Summary

Failure Type(s)	Failure Mode	Related Entities	Related Functions
AI/ML Programming	Faulty base algorithms for AI	Tactical AI (training data)	1.2.1, 1.2.3
AI/ML Programming	Outdated data on enemy tactics	Tactical AI (training data), Adversary (not shown in Figure 31)	1.1, 1.2.1, 1.2.3
AI/ML Programming	Meaningless patterns used for AI/ML	Tactical AI (ML), SPY-1, AWACs, C2BMC	1.1, 1.2.1, 1.2.3
AI/ML Programming	Ineffective engagement recommended by AI	Tactical AI (ML), C2BMC	1.2.2, 1.2.3, 1.2.6
AI/ML Programming, Adversarial Attack	Training data spillage to enemy forces	Tactical AI (training data), Adversary (not shown in Figure 31)	1.1, 1.2.3
AI/ML Programming, Adversarial Attack	Outdated data on enemy forces (weapon impact)	Tactical AI (training data), UAS	1.2.3
Operational	AI failure to provide timely engagement response	Tactical AI (ML), C2BMC	1.2.2, 1.2.3, 1.2.6, 1.3
Operational, AI/ML Programming	Overwriting/Loss of AI/ ML training data	Tactical AI (ML), SPY-1, AWACs, C2BMC	1.1, 1.2.1, 1.2.3
Operational, AI/ML Programming	Prioritization of new data and training data for AI/ ML	Tactical AI (ML), SPY-1, AWACs, C2BMC	1.1, 1.2.1, 1.2.3

D. AREA DEFENSE

1. Scenario Description

This third and final scenario introduces the conflict inherent in defending two assets simultaneously. A mix of assets are defending an airfield, a forward-operating base housing friendly aircraft. PATRIOT provides point defense of the airfield while THAAD supports area defense of the base and local city. A G/ATOR Marine Corps radar delivers air surveillance. U.S. fighters defend the airspace against other hostile fighters and bombers. Several bomber-launched hostile cruise missiles originate from a standoff distance and are not detected by local sensors. Other upgraded early warning radars detect threats in the region and the Strategic AI informs the tactical level AI/ML/BMA of threats in the region. Strategic AI/BMA informs tactical BMA to negate the bomber threats, but the tactical level BMA non-concurs and chooses to address the cruise missile threats.

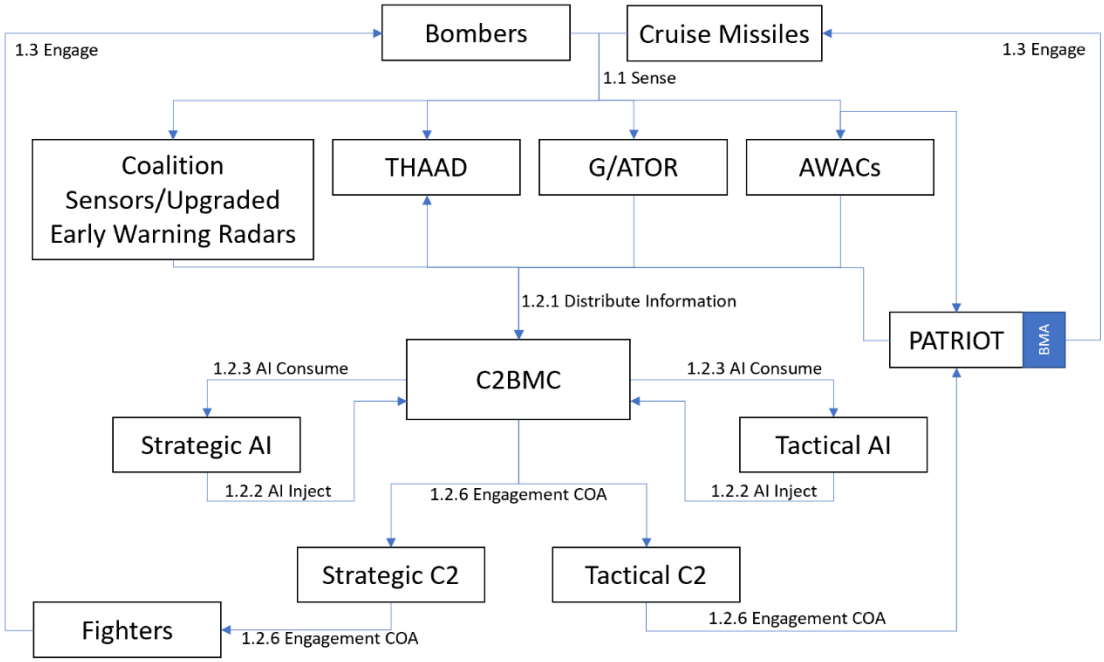

Figure 32. Area Defense Context Diagram

Strategic AI BMA involves an overall encompassing coordinated CONOPS for defending the airfield. The tactical level BMA represents the theater CONOPS that focuses

on a sub-area of interest. The context diagram, in Figure 32, includes an initial look at where the hazard and failure modes are rooted. For this scenario, communication is key to synchronizing the strategic and tactical BMAs. The addition of AI to the strategic BMA introduces different safety risks and associated failure modes.

Bias can be observed when introduced into an AI/ML system at various levels. According to Dietterich and Kong (1995), there is a relative and absolute bias you can introduce into a system. If you consider a series of decision trees, a small decision tree is analogous to tactical level of battle where a much larger decision tree is analogous to strategic/operational battle space. “If these algorithms find a small tree that can correctly classify the training data, then a larger one is not considered” (Dietterich 2005). In this instance, an operational view of the battlefield may not be sufficient to inform the tactical level of battle. In the case of this scenario, there are competing interests that can produce conflicting guidance.

2. Failure Modes and Hazard Analysis

The hazard analysis for this final scenario focuses on one main mishap, which is a successful hostile attack. This scenario is unique since there are multiple incoming hostile threats. The mishap and mishap effect can be caused by one threat or a combination of threats which poses another degree of safety risks. Figure 33 establishes the associated hazards and failure modes. Failure modes are characterized by either a strategic AI BMA failure or a tactical BMA failure, to analyze the strategic vs. theater bias described in the scenario.

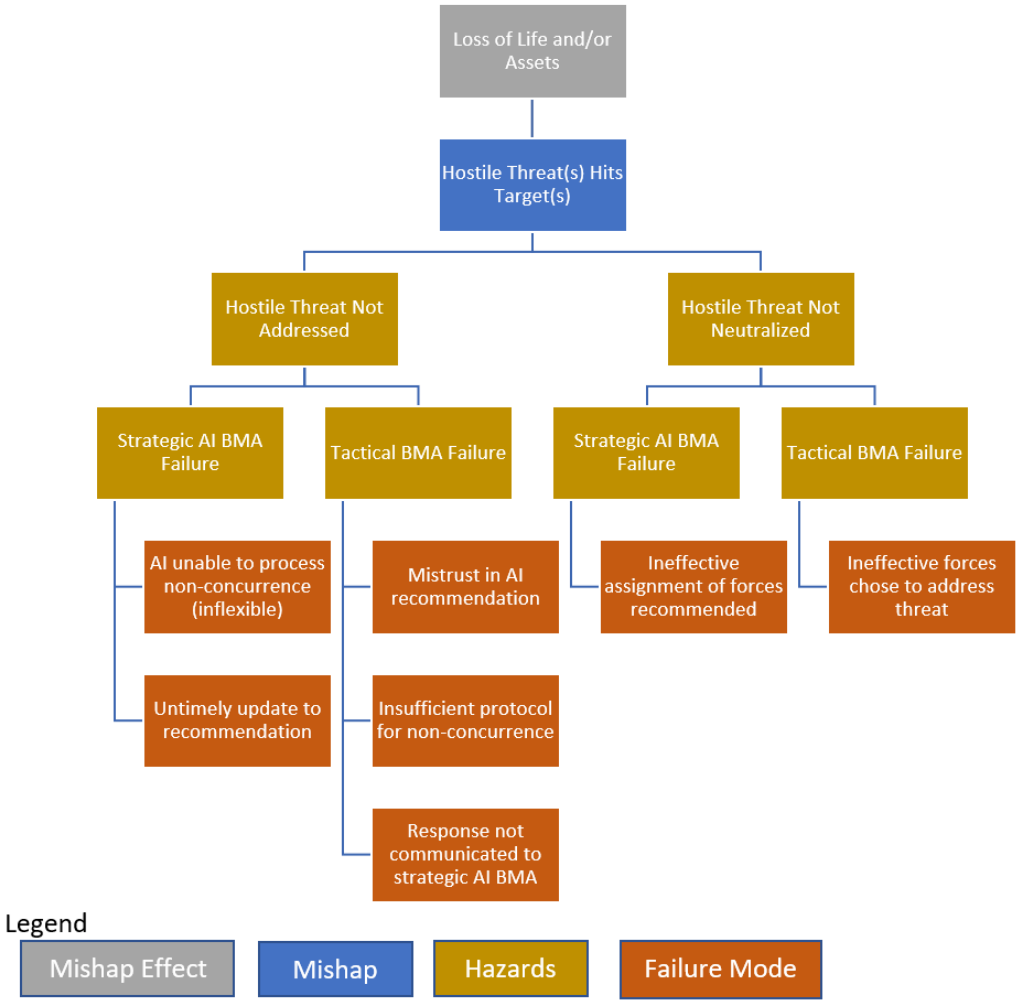

Figure 33. Area Defense Hazard Failure Mode Tree (Strategic vs. Theater Bias)

This scenario involved mostly operational failures, related to overall communication. Communication plays a key role in coordinating engagements from strategic and tactical levels. That coordination needs to include the AI elements and the C2 elements, each operating at the strategic level or the tactical level. The summary in Table 17 also indicated HMI as another main source of error – communication and understanding between human and machine entities.

Table 17. Area Defense (Strategic vs. Theater Bias) Failure Mode Summary

Failure Type(s)	Failure Mode	Related Entities	Related Functions
AI/ML Programming	Strategic AI recommends ineffective assignment of forces	Strategic AI, Fighters, PATRIOT BMA	1.2.2, 1.2.6
Operational	Strategic AI unable to process non-concurrence (inflexible)	Strategic AI, C2BMC, Strategic C2, Tactical C2	1.2.3, 1.2.6
Operational	Untimely update to Strategic AI recommendation	Strategic AI, C2BMC	1.2.2, 1.2.3
Operational, HMI	Mistrust in Strategic AI BMA recommendation	Tactical C2, Tactical AI, Strategic C2	1.2.2, 1.2.3, 1.2.6
Operational, HMI	Insufficient protocol for non-concurrence from Tactical AI/C2	Tactical AI, Tactical C2, C2BMC	1.2.6, 1.3
Operational, HMI	Tactical C2 response to threat not communicated to Strategic AI BMA	Tactical C2, C2BMC, Tactical AI	1.2.2, 1.2.3, 1.2.6
Operational, HMI	Tactical C2 choses ineffective forces to address threat	Tactical AI, Tactical C2, PATRIOT BMA	1.2.6, 1.3

E. SAFETY ANALYSIS FROM AAMD SCENARIOS

The three scenarios reveal several unique AI failures. Artificial intelligence aids in the ultimate COA for AAMD engagement. Therefore, AI failures relate to the cooperation in forming the COA with existing decision makers within the AAMD system of systems, as seen particularly in Scenarios 1 and 3. Scenario 1 covers cooperative failures between AI and WF. In scenario 3, AI fails at the strategic and tactical levels, individually and synchronously. Scenario 2 failure modes deal less with cooperation and more with the programming of AI/ML.

All scenarios consider AI as a source of inefficiency or inaccuracy. The associated hazards depend on the scenario but ultimately attribute to the untimely responses and

ineffective countermeasure choice failure modes. This contributes to the distrust in AI and culture shift needed to accept AI. All identified failure modes hold their own risk and mitigation, as described in the next chapter.

This chapter's hazard analysis provided a few takeaways, revealed from the comparison shown in Table 18. The most common failure types, from the three scenarios, were operational and AI/ML programming. Operational failures mean that system operation or system to system operation caused the failure. The AI/ML programming means incorrect or unintended errors within AI/ML programming caused the failure. All identified failure modes related to communication. Within this system of systems, communication proves to be imperative, especially with AI involvement. Errors between the user and AI played a significant role as well, as seen in the HMI column of Table 18. HMI did not cause the most failures. However, HMI considerations should be a significant consideration during system development. Adversarial attacks introduced unique failure modes but only a few.

Table 18. Failure Mode Comparison from AAMD Scenarios

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
Outdated CONOPs	X					X			1
Incorrect CONOPs/ Training	X					X			1
AI failure to provide timely recommendation	X					X		X	1
AI failure of subsystem error detection	X					X			1
AI failure to provide timely engagement response	X					X	X	X	2

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
Strategic AI unable to process non-concurrence (inflexible)	X					X			3
Untimely update to Strategic AI recommendation	X					X		X	3
Conflicting recommendations	X	X				X			1
AI failure to update time sensitive recommendation	X	X				X		X	1
AI failure of misfire protocol/calculation	X	X			X	X		X	1
Overwriting/Loss of AI/ML training data	X	X			X	X			2
Prioritization of new data and training data for AI/ML	X	X			X	X			2
WF delayed decision (conflicting recommendations)	X	X		X		X			1
WF failure of timely protocols with AI/ ML BMA	X			X		X			1
Limited training	X			X		X			1
Mistrust in Strategic AI BMA recommendation	X			X		X			3
Insufficient protocol for non-concurrence from Tactical AI/C2	X			X		X	X		3
Tactical C2 response to threat not communicated to Strategic AI BMA	X			X		X			3

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
Tactical C2 choses ineffective forces to address threat	X			X		X	X		3
AI failure of countermeasure calculation		X				X			1
AI failure of impact calculation		X				X	X		1
Faulty base algorithms for AI		X				X			2
Outdated data on enemy tactics		X			X	X			2
Meaningless patterns used for AI/ML		X			X	X			2
Ineffective engagement recommended by AI		X				X			2
Strategic AI recommends ineffective assignment of forces		X				X			3
Training data spillage to enemy forces		X	X		X	X			2
Outdated data on enemy forces (weapon impact)		X	X			X			2
WF failure of time sensitive decision				X		X			1

IV. RISK ANALYSIS

In the last chapter, the team described the potential failure modes and hazards associated with a generic AAMD system and three use case threat scenarios. This chapter describes the risk analysis of the potential failure modes of each scenario. The chapter begins with an overview of the risk analysis method. Next, it describes the results of the risk analysis for the generic AAMD system and the three use case scenarios. The chapter ends with a summary of the risk analysis results.

A. RISK ANALYSIS METHOD

NIST Special Publication 800-37 Revision 2 describes the Risk Management Framework (RMF), which provides guidelines to apply the RMF process to information systems and organizations (NIST 2018). According to SP 800-37,

> The RMF includes activities to prepare organizations to execute the framework at appropriate risk management levels. The RMF also promotes near real-time risk management and ongoing information system and common control authorization through the implementation of continuous monitoring processes; provides senior leaders and executives with the necessary information to make efficient, cost-effective, risk management decisions about the systems supporting their missions and business functions; and incorporates security and privacy into the system development life cycle. Executing the RMF tasks links essential risk management processes at the system level to risk management processes at the organization level. In addition, it establishes responsibility and accountability for the controls implemented within an organization's information systems and inherited by those systems.

The RMF is the process that all military branches use to implement privacy and security as well as evaluate the risks present in their information systems. The benefits of RMF are as follows:

- "Provides a repeatable process designed to promote the protection of information and information systems commensurate with risk" (NIST 2018, 2).
- "Emphasizes organization-wide preparation necessary to manage security and privacy risks" (NIST 2018, 2).

- "Facilitates the categorization of information and systems, the selection, implementation, assessment, and monitoring of controls, and the authorization of information systems and common controls" (NIST 2018, 3).
- "Promotes the use of automation for near real-time risk management and ongoing system and control authorization through the implementation of continuous monitoring processes" (NIST 2018, 3).
- "Encourages the use of correct and timely metrics to provide senior leaders and managers with the necessary information to make cost-effective, risk-based decisions for information systems supporting their missions and business functions" (NIST 2018, 3).
- "Facilitates the integration of security and privacy requirements12 and controls into enterprise architecture, SDLC, acquisition processes, and systems engineering processes" (NIST 2018, 3).
- "Connects risk management processes at the organization and mission/ business process levels to risk management processes at the information system level through a senior accountable official for risk management and risk executive (function)" (NIST 2018, 3).
- "Establishes responsibility and accountability for controls implemented within information systems and inherited by those systems" (NIST 2018, 3).

For these reasons, this study uses the RMF process to determine the possible consequences of safety related problems in AI systems used for tactical decision making. This chapter identifies the common risks associated with these systems and describes the risk assessments of this study's three use case scenarios. The risk analysis evaluated the likelihood and impact of the failure modes identified in Chapter III and identified ways to mitigate these risks and mapped the mitigation strategies to the systems engineering life cycle.

1. Risk Determination Process

Risk determinations for each of the failure modes are plotted on a "Risk Diagram, also known as a Risk Matrix, [which] is used to visualize the severity of consequence versus probability" (SPEC Innovations 2021). The failure modes' risks are determined by likelihood of occurrence and impact if the failure were to occur. Levels of likelihood and impact are as follows:

Table 19. Sample Risk Matrix

Levels of	
Likelihood	Impact
Low	Negligible
Medium Low	Minor
Medium	Moderate
Medium High	Serious
High	Critical

After all the risks are plotted, the overall risk for each scenario is determined. These definitions must be analyzed for each organization based on the security posture and nature of the system. The table in Figure 34 shows FIPS 199 potential impacts, which are used to represent the overall risk for Scenarios 1–3.

Table 2: Potential Impact Levels

Potential Impact	Definitions
Low	The potential impact is **low** if—The loss of confidentiality, integrity, or availability could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals.[7] A limited adverse effect means that, for example, the loss of confidentiality, integrity, or availability might: (i) cause a degradation in mission capability to an extent and duration that the organization is able to perform its primary functions, but the effectiveness of the functions is noticeably reduced; (ii) result in minor damage to organizational assets; (iii) result in minor financial loss; or (iv) result in minor harm to individuals.
Moderate	The potential impact is **moderate** if—The loss of confidentiality, integrity, or availability could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals. A serious adverse effect means that, for example, the loss of confidentiality, integrity, or availability might: (i) cause a significant degradation in mission capability to an extent and duration that the organization is able to perform its primary functions, but the effectiveness of the functions is significantly reduced; (ii) result in significant damage to organizational assets; (iii) result in significant financial loss; or (iv) result in significant harm to individuals that does not involve loss of life or serious life threatening injuries.
High	The potential impact is **high** if—The loss of confidentiality, integrity, or availability could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals. A severe or catastrophic adverse effect means that, for example, the loss of confidentiality, integrity, or availability might: (i) cause a severe degradation in or loss of mission capability to an extent and duration that the organization is not able to perform one or more of its primary functions; (ii) result in major damage to organizational assets; (iii) result in major financial loss; or (iv) result in severe or catastrophic harm to individuals involving loss of life or serious life threatening injuries.

Figure 34. Definitions of Potential Impacts. Source: NIST (2008).

Levels of risk are typically determined using quantitative methods and risk determinations can be adjusted based on program risk tolerance. Due to the hypothetical nature of this study, failure mode risks are unquantifiable. Risks are determined to the best of our abilities based on years of experience with the RMF process, risk analysis, performing RMF validations of DOD systems, and deep tactical understanding of the three scenarios. Quantifiable risk analyses are recommended in the future when AI BMAs are developed.

When deciding risk mitigations and where they should be implemented in the engineering life cycle, the DOD 5000 Model shown below in Figure 35 was used. Risk mitigations are assigned to be mitigated in either the Pre-Deployment or Post-Deployment phases of the engineering life cycle. The Pre-Deployment phase consists of concept refinement (CR), technology development (TD), system development and demonstration

(SDD), and production (PD). The Post-Deployment phase includes operations and support (OS). Risk mitigations will need to be analyzed in the future for each AI BMA developed.

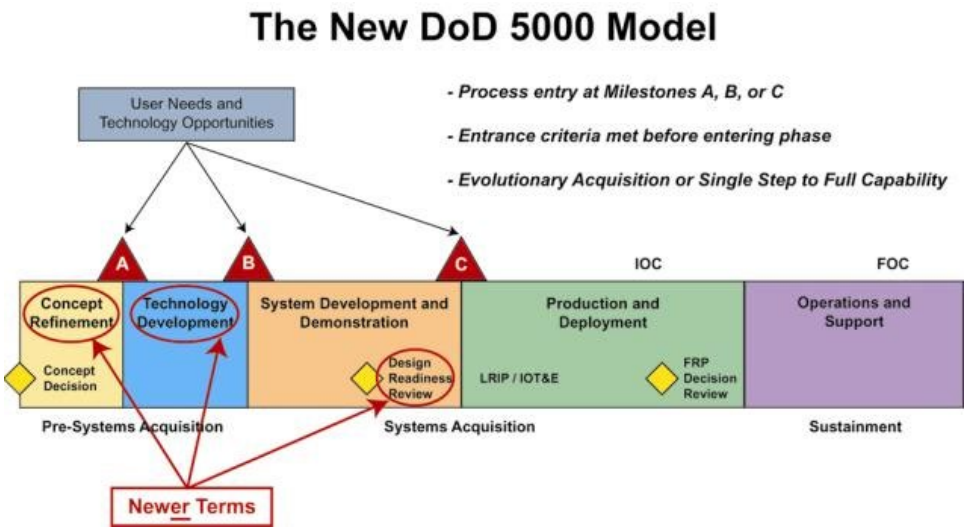

Figure 35. The New DOD 5000 Model. Source: Inflectra (2020)

B. RISK ASSESSMENT

1. Computer AI Systems

Risk levels of common failure modes were determined for general Computer AI systems and evaluated for risk based on the likelihood and impact if the failure mode were to occur using the RMF Process. Failure modes of general computer AI systems were determined based on years of experience performing RMF analysis on numerous computer systems large and small. These failure modes have been visually represented on the Risk Matrix in Figure 36.

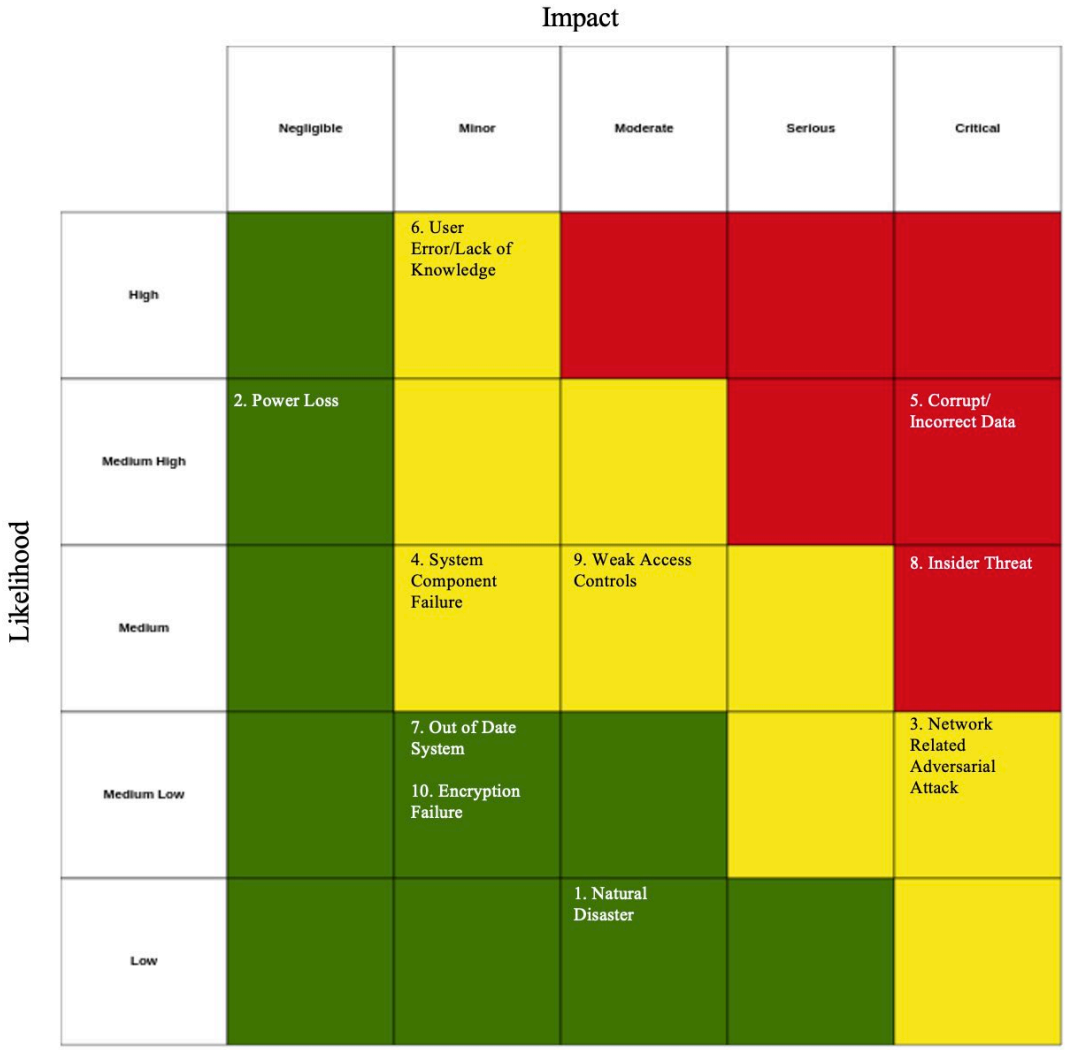

Figure 36. Risk Assessment Matrix - Common System Hazards

Based on the RMF assessment, the overall risk for the common system hazards (explained in further detail below, including the reasoning for the likelihood and impact) led to the overall risk determinations. Although there is no way to completely mitigate risks, risk mitigation recommendations were made to help manage the risks for all failure modes analyzed in Figure 36. It is important to determine when these risk mitigations would need to be developed and implemented within the engineering life cycle, so as to be better prepared for future BMAs leveraging AI and ML for the AMD Mission. Implementation within the Engineering life cycle is also explained in detail below.

1. Natural Disaster

    a. **Risk**: Low

    b. **Risk Determination**: The likelihood of a natural disaster is low. The impact would be moderate because depending on the type of natural disaster a site is susceptible. A Disaster Recovery plan determines how a site handles expected natural disasters.

    c. **Risk Mitigation**: Establish alternate sites and alternate equipment in case of a natural disaster to mitigate this risk.

    d. **Engineering Life Cycle Phases**: During the Pre-Deployment (systems engineering design and development) phase, the Program Management office needs to develop a Disaster Recovery Plan and ensure alternate sites/ alternate equipment are created. During the Post-Deployment (operational) phase, the sites will need to go through tabletop exercises to ensure they are ready in case of disaster, the Disaster Recovery Plan must be reviewed annually and updated as needed, and the alternate sites/equipment must continue to be maintained and updated.

2. Power Loss

    a. **Risk**: Low

    b. **Risk Determination**: The likelihood of power loss is medium high. The impact would be negligible because computer systems should account for this through a power reserve/fault tolerance.

    c. **Risk Mitigation**: This is mitigated through use of uninterruptable power supply (UPS), which ensures the system will continue to run for hours even in the case of lost power.

    d. **Engineering Life Cycle Phases**: During the Pre-deployment phase (concept refinement), planning for the inclusion of UPS to the system is needed. During the Post-Deployment phase (operations) continuous monitoring of the system and the UPS should occur.

3. Network Related Adversarial Attack

    a. **Risk**: Moderate

    b. **Risk Determination**: The likelihood is medium low because most of these BMAs do not connect to the internet and only connect to tactical systems within their boundary. The impact could be critical because if the system were to be compromised, the enemy could control countermeasures and access critical data.

    c. **Risk Mitigation**: Ensuring proper protections are in place, including firewall and network protections, logs auditing all actions, not connecting the system to the internet, having whitelisting programs to protect the system against unauthorized access or modification and having physical access restrictions all greatly reduce this risk.

    d. **Engineering Life Cycle Phases**: During the Pre-Deployment (SDD) phase, network protections need to be implemented into the system. It can also be planned for the system to only connect to a tactical network that does not connect to the Internet (or not, depending on the type of system). Whitelisting can be implemented to ensure no unauthorized executables are run. During the Post-Deployment phase (OS) system admins will need to ensure Solidcore (or other change management device) is running and perform regular log-audits to ensure no unauthorized access occurs.

4. System Component Failure

    a. **Risk**: Moderate

    b. **Risk Determination:** The likelihood of a component failure is medium. The impact would be minor because most systems have fault tolerance.

    c. **Risk Mitigation**: Ensuring fault tolerance and having backup system components readily available reduces this risk.

    d. **Engineering Life Cycle Phase**: During the Pre-Deployment phase (SDD), fault tolerance will need to be implemented in the design of the system.

5. Corrupt/Incorrect Data

    a. **Risk**: High

    b. **Risk Determination**: The likelihood of corrupt or incorrect data is Medium high because if updates are not regularly made, data can be out of date. The impact is critical because if the system has corrupt data, it will not make the correct recommendations or may not be operational.

    c. **Risk Mitigation**: Regular backups and audits mitigate this risk. Additionally, ensuring updates are tested before implementing into the system ensures updates are compatible and that information is intact.

    d. **Engineering Life Cycle Phase**: During the Post-Deployment phase (OS), the system must follow DOD backup policies to ensure all data is backed up in case of the need to restore from previous backup versions. Log audits should be conducted to determine where corrupt data may come from and testing of updates (support) should be done before implementation occurs after the system has been deployed.

6. User Error/Lack of Knowledge or Training

    a. **Risk**: Moderate

    b. **Risk Determination**: The likelihood of user error or lack of knowledge is high; humans make mistakes and training does not always occur. The impact is minor because the AI will be making the recommendations.

    c. **Risk Mitigation**: Establishing standards, regular training and automated processes through the AI mitigates this risk.

    d. **Engineering Life Cycle Phases**: During the Pre-Deployment phase (CR), determine standards and document for training requirements. Ensure the system is user friendly and automate repetitive user tasks (SDD). During the Post-Deployment phase, users will require training updates as the systems evolve (support).

    e.

7. Out of Date System
    a. **Risk**: Low
    b. **Risk Determination**: The likelihood is medium low with a minor impact. Systems are required to be updated regularly to apply the proper patches and updates.
    c. **Risk Mitigation**: Regular updates and update policies help to mitigate this risk.
    d. **Engineering Life Cycle Phases**: During the Pre-Deployment phase (CR), determine and document the system updates policy. During the Post-Deployment phase (OS), the system administrators must abide by the update policy and ensure updates are implemented according to policy frequency and standards.
8. Insider Threat
    a. **Risk**: High
    b. **Risk Determination**: The likelihood is medium with a critical impact. Insider threat is one of the most critical threats because an insider has access to the system.
    c. **Risk Mitigation**: To mitigate this risk, personnel must receive the proper vetting and sign acceptable use policies (AUP). Additionally, regular audits to monitor access and Solidcore (or other whitelisting program) must be implemented to block any escalated privileges or unauthorized modifications.
    d. **Engineering Life Cycle Phases**: During the Pre-Deployment phase (CR), policies for access control and personnel vetting must be determined and documented. The system must be developed to protect against insider threat through the use of whitelisting and privilege limitations by employing account types and passwords (SDD). During the Post-Deployment phase, users must go through the documented vetting process, training and sign

AUPs to access the system (OS). Conduct regular audits to ensure no unauthorized access occurs.

9. Weak Access Controls

    a. **Risk**: Moderate

    b. **Risk Determination**: The likelihood is medium with moderate impact. The DOD requires limiting account privileges and account management policies as well as protecting physical access.

    c. **Risk Mitigation**: Ensure the system follows DOD standards for access control and having physical access protections.

    d. **Engineering Life Cycle Phases**: During the Pre-Deployment phase (CR), the program office must determine access control and physical security policies and document them. The system must be developed to implement access control and physical access protections. During the Post-Deployment phase (OS), the gaining unit must abide by the access control and physical security policies.

10. Encryption Failure

    a. **Risk**: Low

    b. **Risk Determination**: The likelihood is medium low with a minor impact because most systems are programmed to try re-encrypting in the case of an encryption failure.

    c. **Risk Mitigation**: Ensure the system re-encrypts in case of encryption failure.

    d. **Engineering Life Cycle Phase**: During the Pre-Deployment phase (SDD), the system must be developed to re-encrypt in case of encryption failure. During the Post-Deployment phase (OS), the system administrators must ensure regular backups are being performed in case an encryption failure causes the system to malfunction and needs restoration from a previous backup.

The Risk Mitigation Matrix in Table 20 summarizes the risk level of common computer system failure modes, the risk mitigations/recommendations and which part of the engineering life cycle the risk would be addressed.

Table 20. Risk Mitigation – Common System Hazards

Failure Mode	Risk Level	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Common System Hazards			
Natural Disaster	Low	-Alternate sites/Equipment	SDD, OS
Power Loss	Low	-Uninterruptable power supply (UPS)	CR, PD, OS
Network Related Adversarial Attack	Moderate	-Firewall protections -Network protections -Closed systems with no connections to internet	SDD, OS
System Component Failure	Moderate	-Fault Tolerance -Backup system components	SDD
Corrupt/Incorrect Data	High	-Backups -Audits -Testing updates before implementation	OS
User Error/Lack of Knowledge or Training	Moderate	-Training -Standards -Automated processes	CR, SDD, PD, OS
Out of Date System	Low	Regular updates	CR, PD, OS
Insider Threat	High	-Vetting -AUP -Audits -Whitelisting	CR, SDD, PD, OS
Weak Access Controls	Moderate	-Standards -Physical access protection	CR, TD, PD, OS
Encryption Failure	Low	-Re-encryption -Software updates	SDD, OS

The overall risk of computer systems varies from one system to another based on mitigation strategies that are in place, their connections to other systems and the internet, the importance of the system and what they control as well as their availability standards.

The overall risk of an AI BMA will have to be determined once it is designed, and the risks above should be considered.

2. Scenario 1 – Ballistic Missile Defense

Risk levels of each failure mode were determined for Scenario 1, which is a trust deficit between the operator and the AI BMA.

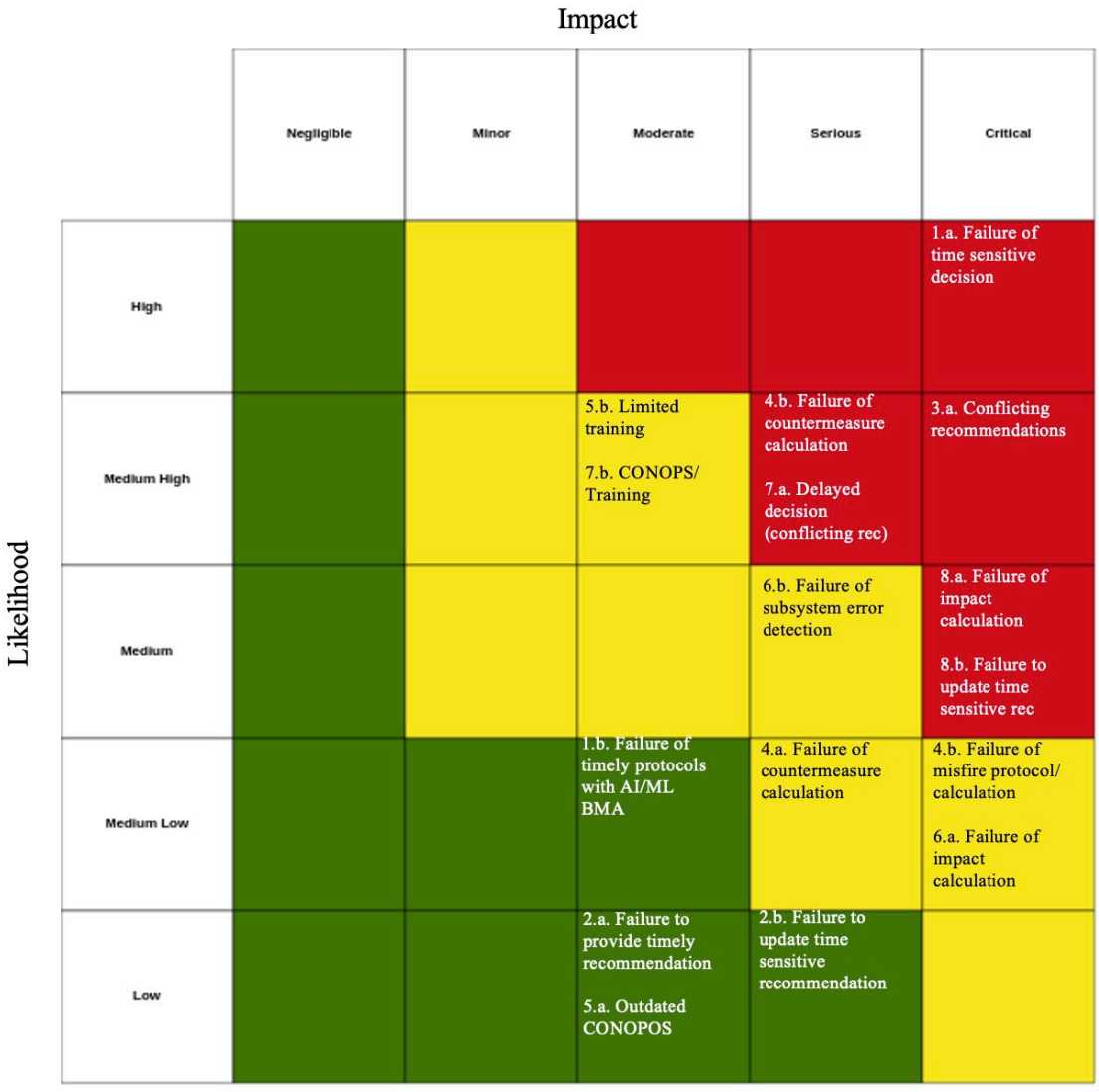

Figure 37. Risk Assessment Matrix – Scenario 1

Based on the assessment, the overall risk for the systems hazards in Scenario 1 are as follows:

1. Ineffective Response Time – User Failure
    a. Failure of time sensitive decision
        i. **Risk:** High
        ii. **Risk Determination:** In this scenario, leadership comes up with a different solution than the AI so the likelihood of not providing a time sensitive decision is high. The impact is critical because if a decision is not made in a timely manner, life and assets are at stake.
        iii. **Risk Mitigation:** Establish user training and standards to ensure leadership makes the most informed decisions. Distributing up to date CONOPS and related AI BMA tactical policies to personnel will help ensure decisions of both the humans and AI correlate. Devising a required reaction time will aid in ensuring action.
        iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase (CR) training and time standards need to be established as well as a CONOPS. Additionally, the system needs to be designed to implement the established polices. During the Post-Deployment phase (OS), the documented standards and CONOPS need to be distributed to system users and they need to receive the proper training. The CONOPs and system policies need to be updated at least annually according to DOD standards.
    b. Failure of timely protocols with AI/ML BMA
        i. **Risk:** Low
        ii. **Risk Determination:** The likelihood is medium low, and the impact is moderate because after a certain amount of time the AI should update to make the best decision based on the most up to date information.

iii. **Risk Mitigation:** Establish user training and standards to ensure the best and prompt decision is made. Regular updates to the CONOPS distributed to personnel and policies programmed in the AI BMA will ensure decisions of both the humans and AI correlate. Establish required reaction time to ensure an action is made.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, training and standards and a required reaction time must be determined and documented (CR, TD). The system must be designed to implement time standards and the proper CONOPS. During the Post-Deployment phase (OS), all training and standards information must be disseminated, and system users must be trained before operating the system.

2. Ineffective Response Time – AI/ML Failure

   a. Failure to provide timely recommendation

      i. **Risk:** Low

      ii. **Risk Determination:** In this case, the AI will still make a timely decision so the likelihood is low, and the impact is moderate only because the system would wait for confirmation from the user which could impact appropriate reactions.

      iii. **Risk Mitigation:** Establish a response time requirement (based on the system) to mitigate this risk. After the required response time passes, the system will update using the latest information to ensure the most correct recommendation is given to the user.

      iv. **Engineering Life Cycle Phase:** For this mitigation, the system must implement a time requirement to wait before updating its recommendation to the user. This is all done during the Pre-Deployment phase (TD).

   b. Failure to update time sensitive recommendation

i. **Risk:** Low

ii. **Risk Determination:** In this situation, the AI makes an initial timely recommendation but the time in which the decision must be made passes and the AI must update the recommendation for engagement since the original is no longer the best option, but the AI fails to notify the user of the updated recommendation. The likelihood is low because the AI should be programmed to continually update and notify of the best recommendation. The impact is serious since an outdated recommendation puts life/assets at risk.

iii. **Risk Mitigation:** A response time standard where the user must respond within the time allowed must be implemented. If this time passes with no response the system needs to update and notify the user of the new best recommendation. Continual updates and alerts help to negate this risk, but time is the biggest factor for risk in this failure mode.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, response time standards must be identified (CR), and the system must be designed and implemented to update to the best recommendation once the allowed time for user response has passed. During the Post-Deployment phase, the system needs to continually update with the latest information (OS).

3. Ineffective Countermeasure – User Failure

    a. Conflicting recommendations

        i. **Risk:** High

        ii. **Risk Determination:** The likelihood is medium high since the user and the system might not be following the same process for

determining the best COA. Impact is critical because if the user does not trust the AI the wrong decision may be made.

iii. **Risk Mitigation:** The CONOPS must be updated at least annually according to DOD standards. Additionally, ensuring the AI BMA is programmed to follow the CONOPS and is updated regularly when CONOPS updates are made will reduce the risk.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase (CR), the CONOPS must be developed and documented, and the system must be designed and implemented (SDD) to follow the CONOPS. During the Post-Deployment stage, the CONOPS must be updated at least annually (or additionally as needed)(OS), and users must be trained regularly and follow the most up to date CONOPS to help ensure similar behavior between the users and the AI (OS).

4. Ineffective Countermeasure – AI/ML Failure

    a. Failure of countermeasure calculation

        i. **Risk:** Moderate

        ii. **Risk Determination:** The likelihood is medium low because the AI BMA will still make the countermeasure calculation based on the information it is given and based on its programming. The impact is serious if the AI BMA was not programmed according to the CONOPS, the calculation could be incorrect, and life/assets would be at risk.

        iii. **Risk Mitigation:** Follow the most up to date CONOPS processes. Account for user's recommendations, then make the best decision based on the data it has and the data given by the user.

        iv. **Engineering Life Cycle Phase:** During the Pre-Deployment phase, the system must be designed and implemented (SDD) to follow the

most up to date CONOPS and account for the user's recommendations. During the Post-Deployment phase, the system administrators would need to implement updates (OS) when updates to the CONOPS are made.

b. Failure of misfire protocol/calculation

i. **Risk:** High

ii. **Risk Determination:** The likelihood is medium high because the AI BMA in this scenario might not be using the same processes as the leadership/user, which also makes the impact critical because life/assets can be a stake.

iii. **Risk Mitigation:** The system must follow the most up to date CONOPS processes. It should also account for the user's recommendations then make the best decision based on the data it has, and the data given by the user.

iv. **Engineering Life Cycle Phase:** During the Pre-Deployment phase (TD), the system must be designed and implemented to take in user recommendations as data to then produce the best recommendation given the collected data and any data the user provides.

5. Lethal Object Selection – User Failure

a. Outdated CONOPS

i. **Risk:** Low

ii. **Risk Determination:** The likelihood is low because the DOD has a requirement for at least annual updates, but the impact would be moderate because if the CONOPS is not up to date, there is moderate risk.

iii. **Risk Mitigation:** Ensuring the CONOPS is updated at least annually will help to mitigate this risk.

iv. **Engineering Life Cycle Phase:** During the Post-Deployment phase, the CONOPS must be updated at least annually or additionally as needed (OS). The system must be updated as the CONOPS is updated (OS).

b. Limited Training

i. **Risk:** Moderate

ii. **Risk Determination:** The likelihood is medium high because training processes are not always followed or implemented and the impact is moderate because with untrained users, there is moderate risk.

iii. **Risk Mitigation:** Ensure an onboarding process is in place for new personnel, ensure they are properly trained and receive annual refresher training / additional training as needed to reduce this risk.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, training and procedures must be determined, developed and documented (CR). During the Post-Deployment phase, users must be trained and onboarded according to policy before using the system (OS). Users must receive annual refresher training or additional training as needed according to policy (OS).

6. Lethal Object Selection – AI/ML Failure

a. Failure of impact calculation

i. **Risk:** Moderate

ii. **Risk Determination:** The likelihood is medium low because the AI BMA will still make the countermeasure calculation based on the information it is given and based on its programming. The impact is critical if the AI BMA was not programmed according to the CONOPS; the calculation could be incorrect, and life/assets would be at risk.

iii. **Risk Mitigation:** The system must follow the most up to date CONOPS processes. It should also account for the user's recommendations, then make the best decision based on the data it has and the data given by the user.

iv. **Engineering Life Cycle Phase:** Before the system is deployed, the system must be designed (CR) to implement the most up to date CONOPS procedures and account for the user's input to make the best recommendation based on collected information and user input. During the Post-Deployment phase, the system must be updated as updates to the CONOPS or algorithms are created (OS).

b. Failure of subsystem error detection

i. **Risk:** Moderate

ii. **Risk Determination:** The likelihood is medium, and the impact is serious if the system is not following the same processes as the user. If it does not detect an error and there is one, it can put life and assets in grave danger.

iii. **Risk Mitigation:** The system must follow the most up to date CONOPS processes. It should also account for the user's recommendations, then make the best decision based on the data it has and the data given by the user and ensure error detection is enabled.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed to implement the most up to date CONOPS procedures and account for the user's input to make the best recommendation based on collected information and user input (SDD). The system must follow the most up to date algorithms to ensure error detection occurs (TD). During the Post-Deployment phase, the system must be updated as updates to the CONOPS or algorithms are created (OS).

7. Location/Timing of Countermeasure Impact – User Failure

    a. Delayed decision (conflicting recommendations)

        i. **Risk:** High

        ii. **Risk Determination:** The likelihood is medium high because the user might not make a timely decision, especially when the AI came up with a different solution. Failure to make a timely decision would make the impact serious since life/assets would be at stake.

        iii. **Risk Mitigation:** Update the recommendation based on the latest information if the user does not respond within a set time standard. Both the user and the AI system must be using the same CONOPS for decisions. Additionally, ensure users are properly trained to make a timely decision.

        iv. **Engineering Life Cycle Phase:** During the Pre-Deployment phase, time standards must be established (CR), and the system must be designed and implemented to update the recommendation if the time threshold for response passes (TD). During the Post-Deployment phase, the system must be updated with new CONOPS polices (OS), and users must be properly trained and be issued the most up to date CONOPS (OS).

    b. CONOPS/Training

        i. **Risk:** Medium

        ii. **Risk Determination:** The likelihood is medium high because updates and training standards are not always followed, and the impact is moderate if the users are not trained or informed correctly.

        iii. **Risk Mitigation:** Ensure the CONOPS is updated at least annually and that users are trained.

        iv. **Engineering Life Cycle Phases:** During the Post-Deployment phase, the CONOPS must be updated at least annually and

disseminated to users of the system. Processes must be updated on the system and updates must be made to user training (support).

8. Location/Timing of Countermeasure Impact – AI/ML Failure

    a. Failure of impact calculation

        i. **Risk:** High

        ii. **Risk Determination:** The likelihood is medium because the system might not be following the same processes as the user when coming up with recommendations. The impact is critical because if the calculation is incorrect, life/assets are at risk.

        iii. **Risk Mitigation:** The system must be following the most up to date CONOPS processes. It must account for users' recommendations, then make the best decision based on the data it has and the data given by the user.

        iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed to take input from the user and determine the best recommendation based on the information collected and user input (CR, TD). During the Post-Deployment phase, the CONOPS must be updated and disseminated to system users, and the system must be updated to follow the most recent CONOPS (OS).

    b. Failure to update time sensitive recommendation

        i. **Risk:** High

        ii. **Risk Determination:** The likelihood is medium because the system might not be following the same processes as the user when coming up with recommendations. The impact is critical because if the recommendation is not updated on time, life/assets are at risk.

        iii. **Risk Mitigation:** The system must be following the most up to date CONOPS processes. It should also account for users'

recommendations, then make the best decision based on the data it has and the data given by the user.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed to take input from the user and determine the best recommendation based on the information collected and user input. During the Post-Deployment phase, the CONOPS must be updated and disseminated to system users, and the system must be updated to follow the most recent CONOPS.

Although the risk levels are high for this scenario, there are mitigations that can be implemented for these risks to be lowered and/or mitigated. See Table 21 for mitigation recommendations.

Table 21. Risk Mitigation Matrix – Scenario 1

Failure Mode	Risk Level	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Ineffective Response Time – User Failure			
Failure of time sensitive decision	High	-User training/standards -Up to date CONOPS -Required reaction time	CR, OS
Failure of timely protocols with AI/ ML BMA	Low	-User training/standards -Up to date CONOPS -Required reaction time	CR, TD, OS
Ineffective Response Time – AI/ML Failure			
Failure to provide timely recommendation	Low	-Response time standards	TD
Failure to update time sensitive recommendation	Low	-Response time standards -Regular updates and alerts	CR, OS
Ineffective Countermeasure – User Failure			
Conflicting recommendations	High	-Annual CONOPS updates -Programming to ensure AI meets CONOPS	CR, SDD, PD OS
Ineffective Countermeasure – AI/ML Failure			

Failure Mode	Risk Level	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Failure of countermeasure calculation	Moderate	-Programming -Allowing analysis of user input	SDD, PD OS
Failure of misfire protocol/ calculation	Moderate	-Programming -Allowing analysis of user input	TD, PD
Lethal Object Selection – User Failure			
Outdated CONOPS	Low	-Annual updates as required by the DOD.	OS
Limited training	Moderate	-On boarding training -Annual refresher training	CR, OS
Lethal Object Selection – AI/ML Failure			
Failure of impact calculation	Moderate	-Programming -Allowing analysis of user input	TD, SDD, PD, OS
Failure of subsystem error detection	Moderate	-Programming -Allowing analysis of user input	TD, SDD, PD OS
Location/Timing of Countermeasure Impact – User Failure			
Delayed decision (conflicting recommendations)	High	-Response time standards -User training	CR, TD, OS
CONOPS/Training	Moderate	-Annual updates -Annual refresher training	OS
Location/Timing of Countermeasure Impact – AI/ML Failure			
Failure of impact calculation	High	-Programming -Allowing analysis of user input	TD, SDD, PD, OS
Failure to update time sensitive recommendation	High	-Programming -Allowing analysis of user input	TD, SDD, PD, OS

To determine the overall risk, all risks were considered, and the average was determined. For this scenario, the Overall risk is Moderate. In Scenario 1- Ballistic Missile Defense, the failure mode with the highest risk is 1.a. Failure of time sensitive decision with an overall high risk and the failures modes with the lowest risks were 2.a. Failure to provide timely recommendation and 5.a. Outdated CONOPS with an overall risk of Low.

Having the highest risk, failure mode 1.a. makes it clear that the same set of standards must be followed by both the users and the AI system. Having up-to-date policies and systems as well as concordance between the AI and the users is crucial to mitigating most of the risks from Scenario 1.

3. Scenario 2 – Ship Self Defense Training Data

Risk levels of each failure mode were determined for Scenario 2 – Ship Self Defense Training Data in which mishaps occur from incorrect training data.

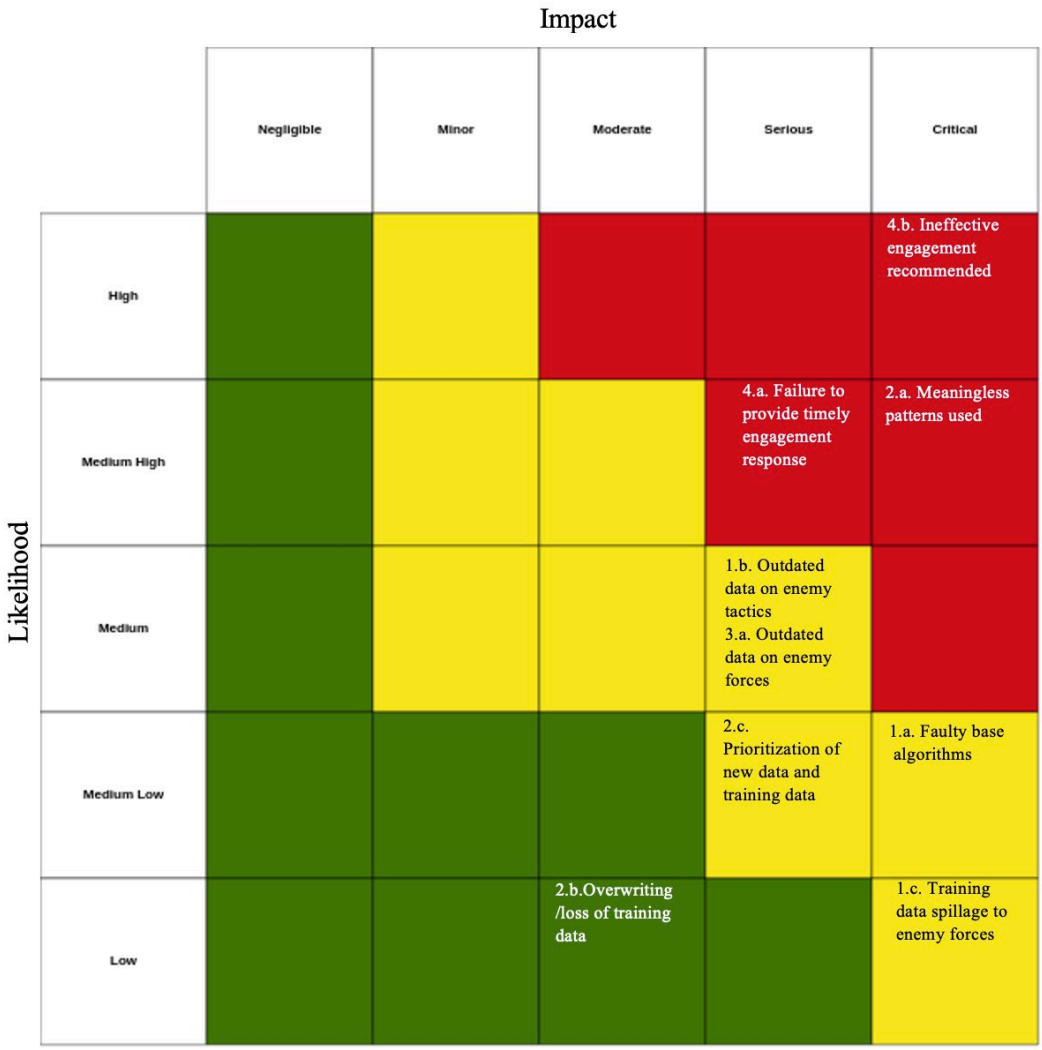

Figure 38. Risk Assessment Matrix – Scenario 2

Based on the assessment, the overall risk for the failure modes in Scenario 2 are as follows:

1. Misidentification - Training Failure
    a. Faulty base algorithms
        i. **Risk:** Moderate
        ii. **Risk Determination:** The likelihood of occurrence is medium low because the system would need to be fielded with the correct and most up to date algorithms in which the system would have to be programmed to prioritize the best algorithm to use for the threat at hand. The impact is critical since life/assets are at stake.
        iii. **Risk Mitigation:** Along with programming the system to use proper algorithms, the algorithms and prioritization must be updated regularly to keep up with current and future threats. Updates must be made at least monthly but have the capability of updating more often depending on when new algorithms or threats are found.
        iv. **Engineering Life Cycle Phases:** Designing and implementing the system to use proper algorithms and prioritization would occur in the Pre-Deployment phase (TD). During Post-Deployment, the system must be updated regularly (OS).
    b. Outdated data on enemy tactics
        i. **Risk:** Moderate
        ii. **Risk Determination:** The likelihood of occurrence is medium because update processes would have to be put into place, and the system would have to be updated regularly in

the field. The impact is serious since life/assets are at stake with outdated information.

iii. **Risk Mitigation:** Updates on enemy tactics must be performed at least monthly or more often when updates emerge.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed to constantly update recommendations based on enemy locations and behaviors (SDD). During the Post-Deployment phase, the system must be updated to use the most up to date information on enemy tactics according to organizational frequency (OS).

c. Training data spillage to enemy forces

i. **Risk:** Moderate

ii. **Risk Determination:** The likelihood of occurrence is low because there would be many protection measures in place preventing disclosure of information to enemies. The impact is critical because in the case of information spillage, life/ assets are at stake.

iii. **Risk Mitigation:** To ensure information does not get in the wrong hands, the system must be encrypted, have firewall rules in place, anti-virus and other software to prevent access/modification of information, and proper access restrictions including the vetting of users/admins and physical access protections.

iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed to implement proper protection against data spillage and to use

whitelisting programs (TD). Training and vetting processes must be determined and documented in this phase as well. During the Post-Deployment phase, users must be trained and vetted to practice good computer security and operational security standards (OS).

2. Misidentification – AI/ML Failure

    a. Meaningless patterns used

        i. **Risk**: High

        ii. **Risk Determination:** The likelihood of occurrence is medium high in this scenario because the system did not have the right data for it to make an informed decision. If updated information were programmed into the system, the likelihood would decrease. The impact is critical because life/assets are at stake.

        iii. **Risk Mitigation:** Proper programming and updates is the best mitigation for this. Additionally, having a proper study done on friendlies would ensure the AI system can identify the proper behavior and determine any discrepancies.

        iv. **Engineering Life Cycle Phases:** During the Pre-Deployment phase, the system must be designed and implemented to use the best and most up to date algorithms to determine recommendations (TD). Studies must be continually conducted on friendly and enemy behaviors and tactics and be provided as information within the AI system (support). The system must be continuously updated as knew studies and information are found, and this would be implemented in the Post-Deployment phase (OS).

        v.

b. Overwriting/Loss of training data

   i. **Risk:** Low

   ii. **Risk Determination:** The likelihood of occurrence is low in this scenario because the system should have backups in place as well as offloading of data onto a separate system for information backups. The impact is moderate because backups should still be in place but overwriting of data can cause unwanted recommendations.

   iii. **Risk Mitigation:** Backups and offloading are mitigations for this, so in case of information loss it can be retrieved from backups. Additionally, no data should be overwritten unless disk space is an issue in which the oldest data would be overwritten first.

   iv. **Engineering Life Cycle Phase:** During the Post-Deployment phase, system admins must ensure backups are being created and offloaded onto a separate system ensuring all data is stored properly and not lost (OS).

c. Prioritization of new data and training data

   i. **Risk:** Moderate

   ii. **Risk Determination:** The likelihood of occurrence is medium low in this scenario because the system should be programmed to prioritize threats appropriately. If the right information were programmed into the system, the likelihood would decrease. The impact is serious because life/assets could be at stake if the wrong prioritization is used.

   iii. **Risk Mitigation:** Proper programming will mitigate this risk to ensure threats are prioritized appropriately. A method for

determining prioritization based on threats must be created and implemented.

iv. **Engineering Life Cycle Phase:** While in Pre-Deployment phase, the system must be designed to implement a proper prioritization strategy (TD). The system must also be updated regularly during the Post-Deployment phase (OS).

3. Ineffective Response – Training Failure
    a. Outdated data on enemy forces
        i. **Risk**: Moderate
        ii. **Risk Determination**: The likelihood of occurrence is medium because update processes would have to be put into place, and the system would have to be updated regularly in the field. The impact is serious since life/assets are at stake with outdated information.
        iii. **Risk Mitigation:** Updates on enemy forces need to be made at least monthly or more often as new updates come about.
        iv. **Engineering Life Cycle Phase:** Updates are made during the Post-Deployment phase (OS).
4. Ineffective Response – AI/ML Failure
    a. Failure to provide timely engagement response
        i. **Risk**: High
        ii. **Risk Determination:** The likelihood of occurrence is medium high in this scenario because the system did not have the right information to make an informed and timely decision. The impact is serious because life/assets are at stake.

iii. **Risk Mitigation:** Proper programming and time standards for response would mitigate this risk. Time for response must update based proximity as well.

iv. **Engineering Life Cycle Phase:** Time standards must be determined (CR) and proper programming of the AI system must be implemented during the Pre-Deployment phase (SDD).

b. Ineffective engagement recommended

i. **Risk**: High

ii. **Risk Determination:** The likelihood of occurrence is high in this scenario because the system did not have the correct data to make a correctly informed decision. If the right information were programmed into the system, the likelihood would decrease. The impact is critical because life/assets are at stake.

iii. **Risk Mitigation:** Proper programming and updates are the best mitigation for this. Additionally, having a proper study conducted on friendlies would ensure the AI system can identify the proper behaviors and determine any discrepancies.

iv. **Engineering Life Cycle Phases:** Proper design (TD) and programming (SDD) must be implemented in the Pre-Deployment phase. Continual studies of behaviors and discrepancies of both friendly and enemy forces must continually be done, and updates would be added during the Post-Deployment phase (OS).

The Risk Mitigation Matrix in Table 22 summarizes the risk level of each failure mode, the risk mitigations/recommendations and which part of the engineering life cycle the risk would be addressed.

Table 22. Risk Mitigation Matrix – Scenario 2

Failure Mode	Risk Level	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Misidentification - Training Failure			
Faulty base algorithms	Moderate	-Ensure proper algorithms used -Ensure most up to date algorithms used	TD, OS
Outdated data on enemy tactics	Moderate	-Updates monthly/as needed	SDD, OS
Training data spillage to enemy forces	Moderate	-Encryption -Firewall -Access restrictions	TD, PD, OS
Misidentification – AI/ML Failure			
Meaningless patterns used	High	-Programming to use proper algorithms	TD, PD, OS
Overwriting/Loss of training data	Low	-Backups -Off loading	OS
Prioritization of new data and training data	Moderate	-Programming to properly prioritize	TD, PD, OS
Ineffective Response – Training Failure			
Outdated data on enemy forces	Moderate	-Updates monthly/as needed	OS
Ineffective Response – AI/ML Failure			
Failure to provide timely engagement response	High	-Response time standards	CR, SDD
Ineffective engagement recommended	High	-Programming of proper algorithms and prioritization	TD, SDD, PD, OS

To determine the overall risk, all risks were considered, and the average was determined. For this scenario, the Overall risk is High. In Scenario 2- Ship Self Defense

Training Data, the failure mode with the highest risk is 4.b., Ineffective engagement recommended with an overall high risk and the failure mode with the lowest risk was 2.b., Overwriting/loss of training data with an overall risk of Low. Having the highest risk, failure mode 4.b makes it clear the importance of the development and design phase of the AI BMA. The system must be programmed with the right information to make the most accurate recommendation. In Scenario 2, much of the risk can be reduced with proper programming, studies of friendly and enemy forces, training, and updates.

4. Scenario 3 – Strategic vs. Theater Bias

Risk levels of each failure mode were determined for Scenario 3 – Strategic vs. Theater Bias in which two assets have conflicting recommendations.

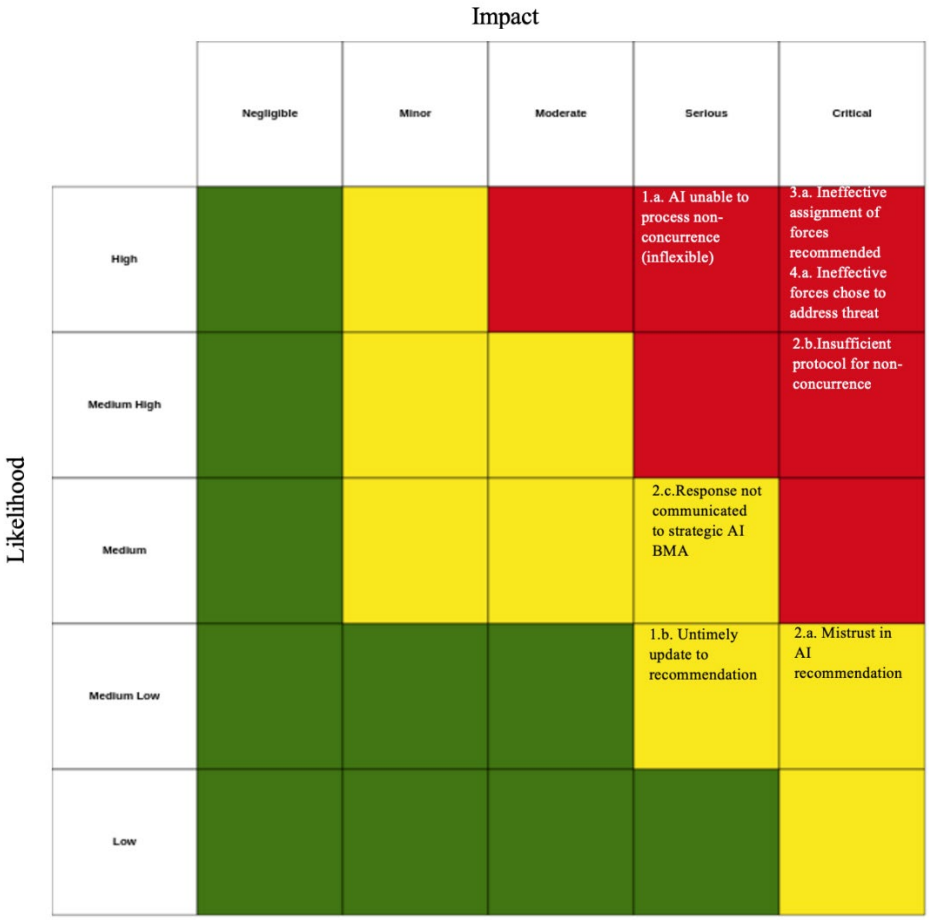

Figure 39. Risk Assessment Matrix – Scenario 3

Based on the assessment, the overall risk for the failure modes in Scenario 3 are as follows:

1. Hostile Threat Not Addressed – Strategic AI BMA Failure
    a. Artificial intelligence unable to process non-concurrence (inflexible)
        i. **Risk**: High
        ii. **Risk Determination**: The likelihood of occurrence is high because in this scenario, the system was not programmed to understand non-concurrence from other systems. The impact is serious because the inability to adjust to other AI systems and decide what the most important threat is puts life/assets at stake.
        iii. **Risk Mitigation**: These systems must be programmed to work with other AI BMA systems and take information from multiple systems to determine the most critical threat/ prioritize actions. This is based on programming before deployment and updates to compatibility post-deployment as other AI systems are deployed. This means these systems must work with other systems from different branches of the military, meaning the programming would go through joint requirements.
        iv. **Engineering Life Cycle Phases**: Proper design and implementation would occur in the Pre-Deployment phase (CR). Studies and compatibility testing of AI BMA systems across all branches would take place in the Pre-Deployment phase as well (CR). These systems must be updated in the Post-Deployment phase to continue to be compatible with new and changing systems and technologies (OS).

b. Untimely update to recommendation

i. **Risk**: Moderate

ii. **Risk Determination**: The likelihood of occurrence is medium low because time standards would be in place to ensure timely recommendations are made. The impact is serious because life/assets are at stake.

iii. **Risk Mitigation**: Ensure the system abides time standards and is programmed to handle multiple recommendations, within the time standards.

iv. **Engineering Life Cycle Phase**: During the Pre-Deployment phase, time standards must be established and documented (CR). Additionally, system design and implementation to handle multiple recommendations would be handled in this phase (SDD).

2. Hostile Threat Not Addressed – Tactical BMA Failure

a. Mistrust in AI recommendation

i. **Risk**: Moderate

ii. **Risk Determination**: The likelihood of occurrence is medium low because the systems would be programmed to work with other BMAs, accounting for all recommendations. The impact is critical because if the correct hostile threat is not addressed, life/assets are at stake.

iii. **Risk Mitigation**: The systems must be programmed to work with other AI BMA systems, account for all recommendations and data found and then make the most informed decision and prioritize neutralizing the biggest threat first.

iv. **Engineering Life Cycle Phase**: A lot of work must be done during the Pre-Deployment phase to ensure the system is designed to work with many recommendations across joint platforms (CR), prioritize and ensure the best overall recommendation is made (TD).

b. Insufficient protocol for non-concurrence

i. **Risk**: High

ii. **Risk Determination**: The likelihood of occurrence is medium high because in this scenario, the AI system was not programmed to take in information from other BMAs to make the correct decision. The impact is critical because if there are no protocols in place, life/assets are at stake.

iii. **Risk Mitigation**: Program the AI BMA systems with the proper algorithms to determine the best course of action based on information from all BMAs they are working with.

iv. **Engineering Life Cycle Phase**: Ensure the system is programmed with proper algorithms during the Pre-Deployment phase (SDD). As updates are made, they must be implemented during the Post-Deployment phase (OS).

c. Response not communicated to strategic AI BMA

i. **Risk**: Moderate

ii. **Risk Determination**: The likelihood of occurrence is medium. The impact is serious because communication between the BMAs is essential for protecting life and assets.

iii. **Risk Mitigation**: Real-time communication must be implemented in these systems as well as the ability to work with multiple AI BMAs.

iv. **Engineering Life Cycle Phases**: Designing the AI systems to communicate with each other across joint platforms must occur in the Pre-Deployment phase CR, TD). A conscious effort to update all communicating systems would be a task for the system administrators during Post-Deployment (OS).

3. Hostile Threat Not Neutralized – Strategic AI BMA Failure

    a. Ineffective assignment of forces recommended

        i. **Risk**: High

        ii. **Risk Determination**: The likelihood of occurrence is high because in this scenario, the system was not programmed to understand non-concurrence from other systems. The impact is critical because the inability to adjust to other AI systems and determine what the most important threat puts life/assets at stake.

        iii. **Risk Mitigation**: These systems must be programmed to work with other AI BMA systems and take information from multiple systems to determine the most critical threat and prioritize actions. This is based on programming before deployment and updates to compatibility post-deployment as other AI systems are deployed. This means these systems must be created to work with other systems from different branches of the military, meaning the programming would go through joint requirements.

        iv. **Engineering Life Cycle Phases**: Designing the AI systems to communicate with each other across joint platforms would occur in the Pre-Deployment phase (CR). A conscious effort to update all communication systems would be a task for the system administrators during Post-Deployment (OS).

4. Hostile Threat Not Neutralized – Tactical BMA Failure

    a. Ineffective forces chosen to address threats

        i. **Risk**: High

        ii. **Risk Determination**: The likelihood of occurrence is high because in this scenario, the system was not programmed to understand non-concurrence from other systems. The impact is critical because the inability to adjust to other AI systems and determine the most important threat puts life/assets at stake.

        iii. **Risk Mitigation**: These systems must be programmed to work with other AI BMA systems and take information from multiple systems to determine the most critical threat and prioritize actions. This is based on programming before deployment and updates to compatibility post-deployment as other AI systems are. deployed. This means these systems must be created to work with other systems from different branches of the military, meaning the programming would go through joint requirements.

        iv. **Engineering Life Cycle Phases**: Designing the AI systems to communicate with each other across joint platforms would occur in the Pre-Deployment phase (CR). A conscious effort to update all communication systems would be a task for the system administrators during Post-Deployment (OS).

The Risk Mitigation Matrix in Table 23 summarizes the risk level of each failure mode, the risk mitigations/recommendations and which part of the engineering life cycle risk addressed is impacted.

Table 23. Risk Mitigation Matrix – Scenario 3

Failure Mode	Risk Determination	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Hostile Threat Not Addressed – Strategic AI BMA Failure			
AI unable to process non-concurrence (inflexible)	High	-Programming to adjust to and work with other AI BMA systems -Updates to compatibility	TD, PD, OS
Untimely update to recommendation	Moderate	-Response time standards -Capable of processing multiple recommendations	CR, PD, SDD
Hostile Threat Not Addressed - Tactical BMA Failure			
Mistrust in AI recommendation	Moderate	-Capable of processing multiple recommendations to prioritize	CR, PD, TD
Insufficient protocol for non-concurrence	High	-Algorithms in place to accommodate multiple recommendations	SDD, OS
Response not communicated to strategic AI BMA	Moderate	-Real-time communication	CR, TD, PD, OS
Hostile Threat Not Neutralized – Strategic AI BMA Failure			
Ineffective assignment of forces recommended	High	-Algorithms in place to accommodate multiple recommendations	TD, PD, OS
Hostile Threat Not Neutralized - Tactical BMA Failure			
Ineffective forces chose to address threats	High	-Algorithms in place to accommodate multiple recommendations	TD, PD, OS

To assess the overall risk, all risks were considered, and the average was determined. For this scenario, the Overall risk is High. In Scenario 3 - Strategic vs. Theater Bias the failure modes with the highest risks are 3.a., Ineffective assignment of forces recommended and 4.a., Ineffective forces chose to address threat. The failure mode with the lowest risk was 1.b., Untimely update to recommendation with an overall risk of Low. Having the highest risks, failure modes 3.a and 4.a make it clear the importance of the development and design phase of the AI BMA. If these systems are going to work with

other AI BMAs throughout the DOD, they must be programmed for compatibility and interoperability. In Scenario 3, much of the risk can be lessened with proper programming, joint efforts for compatibility and updates.

C. RISK ANALYSIS TAKEAWAYS

1. Overall Risk Levels Summary

To summarize our risk analysis, the following tables show all failure modes organized by their risk level, from low to high risk. Organizing the results in this manner focuses on the highest risk failure modes for prioritization when developing AI BMAs.

Table 24. Failure Modes with Overall Low Risk

Failure Mode	Risk Level
Common System Hazards	
Natural Disaster	**Low**
Power Loss	**Low**
Out of Date System	**Low**
Encryption Failure	**Low**
Scenario 1 – Ballistic Missile Defense	
Failure of timely protocols with AI/ML BMA	**Low**
Failure to provide timely recommendation	**Low**
Failure to update time sensitive recommendation	**Low**
Outdated CONOPS	**Low**
Scenario 2 – Ship Shelf Defense Training Data	
Overwriting/Loss of training data	**Low**

Table 25. Failure Modes with Overall Moderate Risk

Failure Mode	Risk Level
Common System Hazards	
Network related adversarial attack	**Moderate**
System component failure	**Moderate**
User error/lack of knowledge	**Moderate**
Weak access controls	**Moderate**
Scenario 1 – Ballistic Missile Defense	
Failure of countermeasure calculation	**Moderate**
Failure of misfire protocol/calculation	**Moderate**
Limited training	**Moderate**
Failure of impact calculation	**Moderate**
Failure of subsystem error detection	**Moderate**
CONOPS/Training	**Moderate**
Scenario 2 – Ship Shelf Defense Training Data	
Faulty base algorithms	**Moderate**
Outdated data on enemy tactics	**Moderate**
Training data spillage to enemy forces	**Moderate**
Prioritization of new data and training data	**Moderate**
Outdated data on enemy forces	**Moderate**
Scenario 3 – Strategic vs. Theater Bias	
Untimely update to recommendation	**Moderate**
Mistrust in AI recommendation	**Moderate**
Response not communicated to strategic AI BMA	**Moderate**

Table 26. Failure Modes with Overall High Risk

Failure Mode	Risk Level
Common System Hazards	
Corrupt/incorrect data	**High**
Insider threat	**High**
Scenario 1 – Ballistic Missile Defense	
Failure of time sensitive decision	**High**
Failure to provide timely recommendation	**High**
Conflicting recommendations	**High**
Delayed decision (conflicting recommendations)	**High**
Failure of impact calculation	**High**
Failure to update time sensitive recommendation	**High**
Scenario 2 – Ship Shelf Defense Training Data	
Meaningless patterns used	**High**
Failure to provide timely engagement response	**High**
Ineffective engagement recommended	**High**
Scenario 3 – Strategic vs. Theater Bias	
AI unable to process non-concurrence (inflexible)	**High**
Insufficient protocol for nonconcurrence	**High**
Ineffective assignment of forces recommended	**High**
Ineffective forces chose to address threat	**High**

2. Risk Mitigation and Engineering Life Cycle Implementation Summary

Through our analysis, we determined risk mitigations for our failure modes and were able to determine when in the systems engineering life cycle they should be addressed and implemented. The tables below organize the failure modes and their risk mitigations based on which phase of the engineering life cycle they fall under. These tables are in order of the engineering life cycle.

Table 27. Risk Mitigations for Failure Modes during the Concept Refinement (CR) Phase

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Common System Hazards	
Power loss	-Uninterruptable power supply (UPS)
User error/lack of knowledge or training	-Standards
Out of date system	-Updating processes
Insider threat	-Processes for vetting, AUPs, audits and whitelisting
Weak Access Controls	-Standards
Scenario 1 – Ballistic Missile Defense	
Failure of time sensitive decision	-User training standards -CONOPS -Required reaction time
Failure of timely protocols with AI/ML BMA	-User training standards -CONOPS -Required reaction time
Failure to update time sensitive recommendation	-Response time standards
Conflicting recommendations	-CONOPS
Limited training	-On boarding training process
Delayed decision (conflicting recommendations)	-Response time standards
Scenario 2 – Ship Self Defense Training Data	
Failure to provide timely engagement response	-Response time standards
Scenario 3 – Strategic vs. Theater Bias	
Untimely update to recommendation	-Response time standards
Mistrust in AI recommendation	-Capable of processing multiple recommendations to prioritize
Response not communicated to strategic AI BMA	-Real-time communication

Table 28. Risk mitigations for Failure Modes during the Technology Development (TD) Phase

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Common System Hazards	
Weak access controls	-Physical access protections
Scenario 1 – Ballistic Missile Defense	
Failure of timely protocols with AI/ML BMA	-Response time timeout functionality
Failure to provide timely recommendation	-Response time timeout functionality
Failure of misfire protocol/calculation	-Programming for the analysis of user input
Failure of impact calculation	-Programming for the analysis of user input
Failure of subsystem error detection	-Programming for the analysis of user input
Delayed decision (conflicting recommendations)	-Programming to implement response time standards
Failure of impact calculation	-Programming for the analysis of user input
Failure to update time sensitive recommendation	-Programming for the analysis of user input
Scenario 2 – Ship Self Defense Training Data	
Faulty base algorithms	-Ensure proper algorithms used
Training data spillage to enemy forces	-Encryption -Firewall -Access restrictions
Meaningless patterns used	-Programming to use proper algorithms
Prioritization of new data and training data	-Programming to properly prioritize
Ineffective engagement recommended	-Programming to properly prioritize
Scenario 3 – Strategic vs. Theater Bias	
AI unable to process non-concurrence (inflexible)	-Programming to adjust and work with other AI BMA systems
Mistrust in AI recommendation	-Capable of processing multiple recommendations to prioritize
Response not communicated to strategic AI BMA	-Real-time communication
Ineffective assignment of forces recommended	-Algorithms in place to accommodate multiple recommendations
Ineffective forces chose to address threats	-Algorithms in place to accommodate multiple recommendations

Table 29. Risk Mitigations for Failure Modes during the System Development and Demonstration (SDD) Phase

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Common System Hazards	
Natural disaster	-Alternate sites/Equipment
Network related adversarial attack	-Firewall protections -Network protections -Closed systems with no connection to the internet
System component failure	-Fault tolerance -Backup system components
User error/lack of knowledge or training	-Training system development -Automated process
Insider threat	-Vetting processes -AUP development -Audits -Whitelisting
Encryption failure	-Re-encryption
Scenario 1 – Ballistic Missile Defense	
Conflicting recommendations	-Programming to ensure AI meets CONOPS
Failure of countermeasure calculation	-Programming to allow analysis of user input
Failure of impact calculation	-Programming for the analysis of user input
Failure of subsystem error detection	-Programming for the analysis of user input
Failure of impact calculation	-Programming for the analysis of user input
Failure to update time sensitive recommendation	-Programming for the analysis of user input
Scenario 2 – Ship Self Defense Training Data	
Outdated data on enemy tactics	-Programming for enemy tactics information
Failure to provide timely engagement response	-Programming for response time
Ineffective engagement recommended	-Programming to properly prioritize
Scenario 3 – Strategic vs. Theater Bias	
Untimely update to recommendation	-Capable of processing multiple recommendations
Insufficient protocol for non-concurrence	-Algorithms in place to accommodate multiple recommendations

Table 30. Risk mitigations for Failure Modes during the Production and Development (PD) Phase

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Common System Hazards	
Power loss	-Uninterruptable power supply (UPS)
User error/lack of knowledge or training	-Standards
Out of date system	-Updating processes
Insider threat	-Processes for vetting, AUPs, audits and whitelisting
Weak Access Controls	-Standards
Scenario 1 – Ballistic Missile Defense	
Conflicting recommendations	-Programming to ensure AI meets CONOPS
Failure of countermeasure calculation	-Programming
Failure of misfire protocol/calculation	-Programming
Failure of impact calculation	-Programming
Failure of subsystem error detection	-Programming
Failure of impact calculation	-Programming
Failure to update time sensitive recommendation	-Programming
Scenario 2 – Ship Self Defense Training Data	
Training data spillage to enemy forces	-Encryption -Firewall -Access restrictions
Meaningless patterns used	-Programming to use proper algorithms
Prioritization of new data and training data	-Programming to properly prioritize
Ineffective engagement recommended	-Programming of proper algorithms
Scenario 3 – Strategic vs. Theater Bias	
AI unable to process non-concurrence (inflexible)	-Programming to adjust to and work with other AI BMA systems
Untimely update to recommendation	-Capable of processing multiple recommendation
Mistrust in AI recommendation	-Capable of processing multiple recommendations
Response not communicated to strategic AI BMA	-Real-time communication
Ineffective assignment of forces recommended	-Algorithms in place to accommodate multiple recommendations
Ineffective forces chose to address threats	- Algorithms in place to accommodate multiple recommendations

Table 31. Risk mitigations for Failure Modes during the Operations and Support (OS) Phase

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Common System Hazards	
Natural disaster	-Alternate sites/equipment upkeep
Power loss	-Uninterruptable power supply (UPS) maintenance
Network related adversarial attack	-Firewall and network protections enabling
Corrupt/incorrect data	-Audits -Backups -Testing updates before implementation
User error/lack of knowledge or training	-Training
Out of date system	-Regular updates
Insider threat	-Vetting -AUP signing -Audits
Weak access controls	-Physical access protection
Encryption failure	-Re-encryption -Software updates
Scenario 1 – Ballistic Missile Defense	
Failure of time sensitive decision	-User training -Updates to CONOPS
Failure of timely protocols with AI/ML BMA	-User training -Updates to CONOPS
Failure to update time sensitive recommendation	-Regular updates and alerts
Conflicting recommendations	-Annual CONOPS updates
Failure of countermeasure calculation	-Updates to analysis
Outdated CONOPS	-Annual updates as required by the DOD
Limited training	-Performing on boarding training -Annual refresher training
Failure of impact calculation	-Updates to analysis
Failure of subsystem error detection	-Updates to analysis
Delayed decision (conflicting recommendations)	-User training
CONOPS/Training	-Annual updates -Annual refresher training
Failure of impact calculation	-Updates to analysis
Failure to update time sensitive recommendation	-Updates to analysis
Scenario 2 – Ship Self Defense Training Data	
Faulty base algorithms	-Ensure most up to date algorithms used

Failure Mode	Risk Mitigations for Engineering Lifecyle Phase
Outdated data on enemy tactics	-Updates monthly/as needed
Training data spillage to enemy forces	-Updates to encryption, firewall, access restrictions
Meaningless patterns used	-Updates to algorithms
Overwriting/Loss of training data	-Backups -Off loading
Prioritization of new data and training data	-Updates
Outdated data on enemy forces	-Updates monthly/as needed
Ineffective engagement recommended	-Updates
Scenario 3 – Strategic vs. Theater Bias	
AI unable to process non-concurrence (inflexible)	-Updates to compatibility
Insufficient protocol for non-concurrence	-Updates to algorithms
Response not communicated to strategic AI BMA	-Real-time communication
Ineffective assignment of forces recommended	-Updates to algorithms
Ineffective forces chose to address threats	-Updates to algorithms

3. Overall Chapter Takeaways

In summary, assessing the three scenarios shows that the risks associated with AI BMAs are high due to the short decision time, tactical nature of the system and that they would be used to protect lives and assets. The overall risk of Scenario 1 – Ballistic Missile Defense is Moderate, for Scenario 2 – Ship Self Defense Training Data is High and for Scenario 3 – Strategic vs. Theater Bias is High. Developing trust in AI systems is difficult since the DOD does not want to place the lives of the warfighter in the hands of an AI computer system.

Artificial intelligence BMAs initially would only recommend the best response to an incoming threat, which will assist the war fighter in making a timely and informed decision. Based on the risk analysis performed, failure modes occurred because of common core software or training data quality issues, where risk can be decreased through various common and new mitigation means. These common issues included user training, policies and documentation, updates to the system and documentation, and the design,

programming and implementation of the AI system, especially consideration of training data sets. If risk mitigations are applied, AI BMAs are possible and will benefit the warfighter, but there are many steps that need to occur before this can happen. Appendix A associates the risk mitigations with the stages of the engineering process. Further study conclusions are addressed in Chapter V.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND PATH FORWARD

This section captures the insights from the team's research, summarizes the failure modes and risks found once AI is introduced into a battle management aid in an AAMD environment, and provides a path forward for future work for AI/ML in the DOD. All scenarios will be revisited to show the common failure modes that can be realized when integrating AI into BMA, especially when the warfighter interfaces with a console operating the systems. A final risk posture highlighting primary culprits of failure modes is provided to show where risks may arise in future systems. Our objectives are re-visited and our approach for analysis discussed to show how the team arrived at our failure modes and risks. Finally, this section offers the potential benefit of this study for future use in the DOD acquisition process.

A. CONCLUSIONS

According to various sources of open-source media, three typical concerns stand out when implementing artificial intelligence and machine learning into systems that have an operator in the loop. Each of the scenarios had a unique issue that put the AAMD mission at risk. Scenario 1 drew out the associated risks with trusting artificial intelligence on the battlefield. Scenario 2 explored reporting errors due to bad training data. Scenario 3 looked at risks as it related to conflicting decisions by two competing AI/ML BMA systems. They are:

- Warfighter Trust Deficit
- Training Data
- Bias

The first is general trust of the system with AI/ML present and is the subject of our first use case scenario. According to Galliott (2018), there is a level of too much trust and too little trust. Galliott describes the encounter of the USS *Vincennes* (CG-49) during the Persian Gulf War. The *Vincennes* was equipped with the latest Aegis Combat System (ACS). It was engaged by small boats while on patrol. During the fight, the system did not

identify an aircraft as a civilian airliner. The ACS categorized the airliner as an enemy fighter aircraft and engaged it as a hostile threat, killing all on board. According to Galliott, "Post-accident reporting and analysis discovered that overconfidence in the abilities of the system, coupled with a poor human-machine interface, prevented those aboard the ship from intervening to avoid the tragedy" (Galliott 2018, 128).

Scenario 1 focused on warfighter trust deficit when defending the United States against an incoming ballistic missile using the Ground Base Midcourse Defense system. We highlighted the possibility of conflicting employment guidance for the weapon system. While warfighters are trained on their CONOPs, the AI/ML recommendation might arrive at a different solution. We found that this conflicting guidance is at a medium likelihood of occurrence with a minor impact to the mission. However, like the incident with the *Vincennes*, the operator can possess an overreliance on the system and miss key inputs that would otherwise lead to the correct course of action in an engagement. General trust in the system can be viewed as having too little confidence as well. Not having the WF trust the system enough or building enough confidence in it would hinder employment of a battle management aid that uses AI/ML.

When it is successfully demonstrated in live fire events, AI/ML inspires and builds confidence. Recently, the Army tested a simple architecture using overhead sensors, an airborne platform, artillery and targets. Project Convergence was executed out of Yuma Proving Grounds by Army Futures Command in September 2020 demonstrating that AI/ML contributed to increased identification, detection, tracking and destruction of incoming aerial threats with a success rate of 98% (Cox 2020).

The second scenario analyzed by the team examined the misidentification of a threat due to inappropriate training data. The destroyer had observed various swarms of UAV/UAS in the vicinity of the ship. The ship observed non-hostile swarms and used the observation as input to the AI/ML. When the swarm became hostile, the AI/ML BMA did not recognize the threat. At a high level, the team assessed the risk of poor training data as medium likelihood with a severe impact.

Training data is used to "teach" machine learning or artificial intelligence to achieve a desired level of operational confidence. Algorithms will learn from the data provided. From this data, the AI/ML code will learn patterns and develop decisions. The adage "garbage in, garbage out" is applicable to this risk in that if poor data is provided to the algorithms, such as dated threat information, the AI/ML would not provide the best decision space to the operator.

Concept drift or a shift in machine learning's goal during its life cycle is applicable in this scenario. Many articles exist that discuss concept drift, or dataset shift. According to machinelearningmastery.com, "Concept drift in machine learning and data mining refers to the change in the relationships between input and output data in the underlying problem over time" (Brownlee 2017) This can have severe consequences on a battlefield. A gradually changing set of code within a ML system must be monitored in order to mitigate risk. As an example, ML could inform the BMA that a threat is within a known adversarial test or demonstration range. While the ML has properly identified the threats that are airborne, it may not recognize that the threat can come from that area and improperly categorize the threat as something that poses danger to a defended area.

The third scenario explored a tactical versus strategic/operational interface for BMA. At a strategic/operational level, policies are set in place to direct order of battle elements or ensure pre-planned responses are met. Resources are identified and courses of actions are written to define the missions for air superiority, cyberspace superiority, and space superiority (Department of the Air Force 2015). There is 'left of launch' planning involved that helps senior leaders understand posturing. At a tactical level the individual battles and engagements are fought. In this scenario, systems that had the responsibility to defend a specific location received notification that an inbound threat was entering their area of responsibility. At the strategic level, AI/ML made recommendations based on operational considerations. However, strategic AI/ML provided the warfighter with input specific to the systems defending the airfield. Not all considerations were made at the tactical level and the AI/ML arrived at the option that best served the strategic level. The bias demonstrated in this scenario is a real possibility with multiple systems using AI/ML.

While the consequences of this are serious, the team assessed the probability of occurrence as low.

The team assessed three fundamental issues that pose risks using AL/ML in a BMA system. These risks are at the root of the scenarios analyzed in Chapters III and IV and shown in Table 32. We looked at how potential warfighters' trust deficit can impact real world operations. From a technical perspective, use of AI/ML can increase decision space and better inform the warfighter. However, trust in AI/ML may drive the warfighter to use existing CONOPs and TTPs. Overreliance can lead to detrimental losses. We've assessed that AI/ML Warfighter Trust Deficit as a 2 x 4 risk (medium-low probability with severe consequences) mitigated through rigorous validation and verification (V&V) and training.

As part of the development of an AI/ML BMA, training data will be required so that the system can learn what it is intended for. As mentioned previously, bad training data will lead to bad results. We assess the risk of poor training data as a 3 x 4 (medium likelihood with severe consequences) mitigated through continuous monitoring and evaluation of the system.

Finally, bias and focus drift are a risk to battle management aids that will use AI/ML. As highlighted in our research, AI/ML may not consider a full decision tree if it arrives at a satisfactory conclusion, removing options to the warfighter that may better suit the situation. We assess bias and focus drift as a 4 x 4 (medium-high likelihood with severe consequences) mitigated through continuous monitoring of output and extensive use in wargames and exercises in order to capture bias early in development.

Table 32. Final Risk Posture

Severity					
5					
4				R3	
3				R2	
2				R1	
1					
	1	2	3	4	5
	Probability of Occurrence				
R1	Warfighter Trust (Scenario 1)				
R2	Training Data (Scenario 2)				
R3	Bias / Focus Drift (Scenario 3)				

Most risks identified in Chapter IV can be mitigated during the Pre-Deployment phase of development for a BMA system. Any Post-Deployment risks that are realized will likely take substantial work to mitigate. As in a software intensive program office, a robust V&V process should be implemented in order to validate that the BMA does not realize any of these risks listed above. An extensive continuous development and continuous delivery approach should be used in order to keep pace with the amount of training data required for AI/ML in BMA for an AAMD mission. Continuous monitoring of BMA systems at any level will be needed once operationally deployed to look for failure modes and ensure risk mitigations are in place.

On the subject of policy, the Department of Defense published DOD Directive 3000.09 that provides initial guidance on development and use of autonomous and semi-autonomous weapon systems. It sets the standard on application of lethal or non-lethal force by autonomous systems. The policy states “4.a. Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force” (Department of Defense 2017). The directive goes on to instruct services that before a system is deployed, it must

demonstrate the capability to allow commander and operators to exercise appropriate levels of human judgement before force is applied. The operational scenario involving an Aegis ship under attack by a fighter with pilots is an example where a fully autonomous system driven by AI/ML could violate policy.

B. CONTRIBUTIONS

The team set out to address the safety risks related to the development and implementation of future BMA that use AI/ML as part of the tactical software. All phases of the acquisition cycle were considered, along with known relevant employment of currently fielded systems. The following will review the goals of this report and the approach the systems engineering team took to achieve them.

1. What are the safety risks related to the deployment of AI systems that support future automated tactical decision and mission planning aids?

Our approach began by developing a generic kill chain that can be applied to AAMD systems. The team created an Innoslate model that captured functions and the decomposed subfunctions to map out where possible failure modes may be present. Through our literature review, we looked at issues that the community was facing with the use of AI/ML in DOD and industry systems. To further explore and draw out potential failure modes, the team developed three unique scenarios meant to provide the environment where a failure could be realized. An Operational View –1 was generated to visualize the AI/ML in BMA systems in the context of a real world setting and then used to generate a set of use cases. A second OV-1 was generated depicting a tactical scenario where a local operator relied on the onboard AI/ML BMA system. Scenario 1 looked at possible mistrust of a BMA using AI/ML in the tactical software. Scenario 2 explored a lone ship not understanding what it is observing as a swarm of otherwise harmless UAVs approach the destroyer. Scenario 3 looked at contributing factors to conflicting guidance from AI/ML systems at an operational level vs its tactical counterpart.

Table 33 is a summary roll up of failure modes the team identified across the scenarios, illustrating the common failure types and related functions.

Table 33. Scenario Failure Mode Comparison

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
Outdated CONOPs	X					X			1
Incorrect CONOPs/ Training	X					X			1
AI failure to provide timely recommendation	X					X		X	1
AI failure of subsystem error detection	X					X			1
AI failure to provide timely engagement response	X					X	X	X	2
Strategic AI unable to process non-concurrence (inflexible)	X					X			3
Untimely update to Strategic AI recommendation	X					X		X	3
Conflicting recommendations	X	X				X			1
AI failure to update time sensitive recommendation	X	X				X		X	1
AI failure of misfire protocol/calculation	X	X			X	X		X	1
Overwriting/Loss of AI/ML training data	X	X			X	X			2
Prioritization of new data and training data for AI/ML	X	X			X	X			2
WF delayed decision (conflicting recommendations)	X	X		X		X			1

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
WF failure of timely protocols with AI/ML BMA	X			X		X			1
Limited training	X			X		X			1
Mistrust in Strategic AI BMA recommendation	X			X		X			3
Insufficient protocol for non-concurrence from Tactical AI/C2	X			X		X	X		3
Tactical C2 response to threat not communicated to Strategic AI BMA	X			X		X			3
Tactical C2 choses ineffective forces to address threat	X			X		X	X		3
AI failure of countermeasure calculation		X				X			1
AI failure of impact calculation		X				X	X		1
Faulty base algorithms for AI		X				X			2
Outdated data on enemy tactics		X			X	X			2
Meaningless patterns used for AI/ML		X			X	X			2
Ineffective engagement recommended by AI		X				X			2
Strategic AI recommends ineffective assignment of forces		X				X			3

Failure Mode	Failure Type(s)				Related Function(s)				Scenario
	Operational	AI/ML Programming	Adversarial Attack	HMI	Sense	Communicate	Engage	Kill Assessment	
Training data spillage to enemy forces		X	X		X	X			2
Outdated data on enemy forces (weapon impact)		X	X			X			2
WF failure of time sensitive decision				X		X			1

2. What are the possible consequences of safety related problems in AI systems used in tactical decision making?

By using the Risk Analysis Framework, the team determined that there are many possible consequences including loss of life, loss of assets, compromised systems and information, untrained personnel, out-of-date system information, physical security risks, and more. It was determined that Scenario 1 had an overall risk of Moderate, while Scenarios 2 and 3 had an overall risk of High. The safety related problems were all analyzed as worst-case scenarios based on the likelihood and the impact of occurrence. An important finding within our research was that these risks can be reduced or mitigated through conscious efforts to ensure quality research, data, design, programming, documentation, operational support and sustainment is implemented throughout the engineering life cycle. Security protections that help to mitigate the risks found in this study are summarized in Appendix A, and the phase of the engineering life cycle that they should be implemented are summarized in Appendix B. The team highly recommends that these protections are considered and implemented when AI BMAs are being developed. Through the implementation of security protections, consequences will be less likely and less severe.

C. POTENTIAL BENEFITS

As threats change and our systems roll off the assembly line, understanding risks and failure modes associated with artificial intelligence early in the program life cycle is critical to fielding an effective system. The AAMD domain will only increase as adversarial weapons looking to saturate sensors or overwhelm defense systems. Ensuring that we can match or outpace the threat should be the focus of the program offices once AAMD systems are fielded. It is important to address these risks and failure modes early in a life cycle to maximize focus on how to defeat the merging threats. Any system or BMA programs that realize the risks identified in this report after the system is fielded will see an increase in performance risk. This report will reduce overall program risk (cost, schedule and performance) to BMA programs that will incorporate AI/ML into tactical code if considered early in a program life cycle.

D. PATH FORWARD

Having identified an initial risk posture of AI/ML in BMA systems, the team strongly recommends investigating how AI/ML will be used in BMA systems. Understanding the specific and intended uses will help mitigate risks associated with biases and will foster system confidence.

(1) Implementation of AI/ML at a Tactical Level vs. Operational Level

The team recommends a study on how to best implement AI/ML in the command structure. As AI/ML is introduced to various systems, how an engagement is consummated is very important. The engagement decision tree could be executed at an operational level however, not all information may be considered or available at the operational level.

(2) Use of AI/ML to Gather Data on Threats

Another recommendation is to perform a study on the potential use of AI/ML to gather performance data on existing or new threats. This study should compare AI/ML data collected versus intelligence community findings. The study should look at how AI/ML can contribute to defense design planning based on new threats, or changes in adversarial use of existing threats.

(3) Verification and Validation of AI/ML Intensive Systems

The team recommends a DOD level directive on how V&V will be managed and implemented for systems that will use AI/ML. This directive should include continuous monitoring of AI/ML software and testing of new code.

(4) Reliability for AI/ML Intensive Programs

The team recommends a service level and DOD level reliability study for AI/ML battle management aids. This study should address effectiveness of information used by the warfighter and levels of confidence of data being presented in light of new information.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A

This Appendix summarizes the risk mitigations that must be implemented when building an AI BMA system. Risk Mitigations are organized based on how many risks they would negate/lessen and then alphabetically.

Most Common Mitigations and when to implement in the Engineering Life cycle

Risk Mitigations	Risks Mitigated	Engineering Life Cycle Stages
Programming	-Conflicting recommendations -Failure of countermeasure calculation -Failure of misfire protocol/ calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation -Faulty base algorithms -Meaningless patterns used -Prioritization of new data and training data -Ineffective engagement recommended -AI unable to process non-concurrence -Untimely update to recommendation -Mistrust in AI recommendation -Insufficient protocol for non-concurrence -Response not communicated to strategic AI BMA	Technology Development (TD) System Development and Demonstration (SDD) Production and Development (PD)

Risk Mitigations	Risks Mitigated	Engineering Life Cycle Stages
	-Ineffective assignment of forces recommended -Ineffective forces chose to address threats	
Standards/Documentation (including the updates)	-User error/lack of knowledge -Weak access controls -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Failure to provide timely recommendation -Failure to update time sensitive recommendation -Conflicting recommendations -Outdated CONOPS -Delayed decision (conflicting recommendations) -CONOPS Training -Failure to provide timely engagement response -Untimely update to recommendation	Concept Refinement (CR) Operations and Support (OS)
Updates	-Out of date system -Encryption failure -Failure to update time sensitive recommendation -CONOPS/Training -Faulty base algorithms -Outdated data on enemy tactics -Outdated data on enemy forces -AI unable to process non-concurrence	Operations and Support (OS)
Allowing analysis of user input	-Failure of countermeasure calculation -Failure of misfire protocol/ calculation	Technology Development (TD) System Development and Demonstration (SDD)

Risk Mitigations	Risks Mitigated	Engineering Life Cycle Stages
	-Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation	Production and Development (PD)
Training (including annual refresher training)	-User error/lack of knowledge -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Limited training -Delayed decision (conflicting recommendations) -CONOPS/Training	Concept Refinement (CR) Operations and Support (OS)
Personnel Vetting	-Insider threat -Limited training -Training data spillage to enemy forces	Operations and Support (OS)
Audits	-Corrupt/incorrect data -Insider threat	Operations and Support (OS)
Backups	-Corrupt/incorrect data -Overwriting/loss of training data	Operations and Support (OS)
Firewall protections/ Network protections	-Network related adversarial attack -Training data spillage to enemy forces	Technology Development (TD) System Development and Demonstration (SDD) Production and Development (PD) Operations and Support (OS)
Acceptable Use Policy	-Insider threat	Operations and Support (OS)
Alternate sites/Equipment	Natural disaster	Concept Refinement (CR) Operations and Support (OS)
Automated Processes	-User error/lack of knowledge	System Development and Demonstration (SDD)

Risk Mitigations	Risks Mitigated	Engineering Life Cycle Stages
Backup Components	-System component failure	Operations and Support (OS)
Encryption	-Training data spillage to enemy forces	Operations and Support (OS)
Fault Tolerance	-System component failure	Operations and Support (OS)
Offloading information onto backup system	-Training data spillage to enemy forces	Operations and Support (OS)
Physical access protections	-Weak access controls	Concept Refinement (CR) Operations and Support (OS)
Re-encryption	-Encryption failure	Operations and Support (OS)
Testing before implementation	-Corrupt/incorrect data	Operations and Support (OS)
Uninterruptable power supply (UPS)	-Power loss	Concept Refinement (CR) Production and Development (PD) Operations and Support (OS)
Whitelisting	-Insider threat	Concept Refinement (CR) System Development and Demonstration (SDD)

APPENDIX B

This Appendix summarizes the risk mitigations that must be implemented when building an AI BMA system. Risk Mitigations are organized based on how many risks they would negate/lessen.

Most Common Mitigations in Concept Refinement Phase

Risk Mitigations	Risks Mitigated
Standards/Documentation (including the updates)	-User error/lack of knowledge -Weak access controls -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Failure to provide timely recommendation -Failure to update time sensitive recommendation -Conflicting recommendations -Outdated CONOPS -Delayed decision (conflicting recommendations) -CONOPS Training -Failure to provide timely engagement response -Untimely update to recommendation
Training (including annual refresher training)	-User error/lack of knowledge -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Limited training -Delayed decision (conflicting recommendations) -CONOPS/Training
Alternate sites/Equipment	-Natural disaster
Physical access protections	-Weak access controls
Uninterruptable power supply (UPS)	-Power loss
Whitelisting	-Insider threat

Most Common Mitigations in Technology Development Phase

Risk Mitigations	Risks Mitigated
Programming	-Conflicting recommendations -Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation

Risk Mitigations	Risks Mitigated
	-Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation -Faulty base algorithms -Meaningless patterns used -Prioritization of new data and training data -Ineffective engagement recommended -AI unable to process non-concurrence -Untimely update to recommendation -Mistrust in AI recommendation -Insufficient protocol for non-concurrence -Response not communicated to strategic AI BMA -Ineffective assignment of forces recommended -Ineffective forces chose to address threats
Allowing analysis of user input	-Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation
Firewall protections/ Network protections	-Network related adversarial attack -Training data spillage to enemy forces

Most Common Mitigations in System Development and Demonstration Phase

Risk Mitigations	Risks Mitigated
Programming	-Conflicting recommendations -Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation -Faulty base algorithms -Meaningless patterns used -Prioritization of new data and training data -Ineffective engagement recommended -AI unable to process non-concurrence -Untimely update to recommendation -Mistrust in AI recommendation -Insufficient protocol for non-concurrence -Response not communicated to strategic AI BMA -Ineffective assignment of forces recommended

Risk Mitigations	Risks Mitigated
	-Ineffective forces chose to address threats
Allowing analysis of user input	-Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation
Firewall protections/ Network protections	-Network related adversarial attack -Training data spillage to enemy forces
Automated Processes	-User error/lack of knowledge
Whitelisting	-Insider threat

Most Common Mitigations in Production and Deployment Phase

Risk Mitigations	Risks Mitigated
Programming	-Conflicting recommendations -Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation -Faulty base algorithms -Meaningless patterns used -Prioritization of new data and training data -Ineffective engagement recommended -AI unable to process non-concurrence -Untimely update to recommendation -Mistrust in AI recommendation -Insufficient protocol for non-concurrence -Response not communicated to strategic AI BMA -Ineffective assignment of forces recommended -Ineffective forces chose to address threats
Allowing analysis of user input	-Failure of countermeasure calculation -Failure of misfire protocol/calculation -Failure of impact calculation -Failure of subsystem error detection -Failure of impact calculation -Failure to update time sensitive recommendation
Firewall protections/ Network protections	-Network related adversarial attack -Training data spillage to enemy forces

Risk Mitigations	Risks Mitigated
Uninterruptable power supply (UPS)	-Power loss

Most Common Mitigations in Operations and Support Phase

Risk Mitigations	Risks Mitigated
Standards/Documentation (including the updates)	-User error/lack of knowledge -Weak access controls -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Failure to provide timely recommendation -Failure to update time sensitive recommendation -Conflicting recommendations -Outdated CONOPS -Delayed decision (conflicting recommendations) -CONOPS Training -Failure to provide timely engagement response -Untimely update to recommendation
Updates	-Out of date system -Encryption failure -Failure to update time sensitive recommendation -CONOPS/Training -Faulty base algorithms -Outdated data on enemy tactics -Outdated data on enemy forces -AI unable to process non-concurrence
Training (including annual refresher training)	-User error/lack of knowledge -Failure of time sensitive decision -Failure of timely protocols with AI/ML BMA -Limited training -Delayed decision (conflicting recommendations) -CONOPS/Training
Personnel Vetting	-Insider threat -Limited training -Training data spillage to enemy forces
Audits	-Corrupt/incorrect data -Insider threat
Backups	-Corrupt/incorrect data -Overwriting/loss of training data
Firewall protections/ Network protections	-Network related adversarial attack -Training data spillage to enemy forces
Acceptable Use Policy	-Insider threat

Risk Mitigations	Risks Mitigated
Alternate sites/Equipment	Natural disaster
Backup Components	-System component failure
Encryption	-Training data spillage to enemy forces
Fault Tolerance	-System component failure
Offloading information onto backup system	-Training data spillage to enemy forces
Physical access protections	-Weak access controls
Re-encryption	-Encryption failure
Testing before implementation	-Corrupt/incorrect data
Uninterruptable power supply (UPS)	-Power loss
Whitelisting	-Insider threat

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Allen, Greg C. 2020. *Understanding AI Technology.* Joint Artificial Intelligence Center (JAIC) Department of Defense. https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf.

Brownlee, Jason. 2017. "A Gentle Introduction to Concept Drift in Machine Learning." *Machine Learning Mastery.* December 15. Accessed 2021. https://machinelearningmastery.com/gentle-introduction-concept-drift-machine-learning/.

Cox, Matthew. 2020. "Army's New Target Tracking System Aims to Quicken Artillery Kills." *Military.com.* September 20. Accessed 2021. https://www.military.com/daily-news/2020/09/20/armys-new-target-tracking-system-aims-quicken-artillery-kills.html.

Deep AI, Inc. 2021. "Machine Learning." *DeepAI.* Accessed 2021. https://deepai.org/machine-learning-glossary-and-terms/machine-learning.

Department of Defense. 2017. "Autonomy in Weapon Systems." *DOD Directive 3000.09.* Department of Defense, May.

Department of the Air Force. 2015. *Levels of War.* Doctrinal Publication, Maxwell Air Force Base, AL: United States Air Force.

Dietterich, Thomas G., and Eun Bae Kong. 1995. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms.* Technical Report, Corvalis, OR: Oregon State University.

DOD. 2012. "Department of Defense Standard Practice: System Safety." Military Standard, Department of Defense.

———. 2018. "Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to advance Our Security and Prosperity." Washington, D.C.

Galliott, Jai. 2018. "The Soldier's Tolerance for Autonomous Systems." *Paladyn, Journal of Behavioral Robotics* 124–136.

Grooms, Geoffrey B. 2019. *Artificial Intelligence Applications for Automated Battle Management Aids in Future Military Endeavors.* Technical Report, Monterey, CA: Naval Postgraduate School.

Hao, Karen. 2019. “Technology Review.” *MIT Technology Review.* November 11. Accessed 2021. https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/.

Inflectra. 2020. “Systems Development With DOD 5000.” *Inflectra.* February 7. Accessed 2021. https://www.inflectra.com/ideas/whitepaper/systems-development-with-dod-5000.aspx.

Johnson, Bonnie. 2021a. “Artificial Intelligence Systems: Unique Challenges for Defense Applications.” *Proceedings of the Eighteenth Annual Acquisition Research Symposium.*

———. 2021b. “Research Projects: Complexity, Artificial Intelligence, Directed Energy Warfare & Systems Engineering for AI Systems.” Presentation, Monterey, CA: Naval Postgraduate School.

———. 2021c. “Metacognition for Artificial Intelligence in System Safety.” Presented at the 16th International Conference on Systems (ICONS) Conference, April 2021.

Johnson, Bonnie, and William A. Treadway. 2019. “Artificial Intelligence — An Enabler of Naval Tactical Decision Superiority.” *AI Magazine*, April: 63–78.

Joint Chiefs of Staff. 2017. *Countering Air and Missile Threats* . Accessed 2021. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_01_pa.pdf .

Joint Services – Software Safety Authorities. 2016. *Implementation Process and Tasks Supporting MIL-STD-882E.* Guidebook, Redstone Arsenal, AL: Joint Services – Software Safety Authorities.

Kaempf, G., S. Wolf, and T. Miller. 1993. “Decision Making in the AEGIS Combat Information Center.” *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* 1107–1111.

Marr, Bernard. 2021. *Bernard Marr & Co.* Accessed 2021. https://bernardmarr.com/the-key-definitions-of-artificial-intelligence-ai.

McCarthy, John. 2007. *What is Artificial Intelligence?* Stanford University, Computer Science Department, Stanford, CA.

McKendrick, Kathleen. 2017. *The Application of Artificial Intelligence in Operations Planning.* Point Paper, Ankara, Turkey: NATO Science and Technology Organization.

Miller, S., and B. Nagy. 2021. “Interdependence Analysis for Artificial Intelligence System Safety.” *Acquisition Research Symposium.*

Naval Sea Systems Command. 2021. *NOSSA - Naval Ordnance Safety and Security Activity.* Accessed 2021. https://www.navsea.navy.mil/Home/NOSSA/.

NIST. 2018. *SP 800-37, Revision 2.* Standards Document, Gaithersburg, MD: National Institue of Standards and Technology.

NIST. 2008. *SP 800–60 Volume I Revision 1.* Standards Document, Gaithersburg, MD: National Institue of Standards and Technology.

Schade, U., and M. R. Hieb. 2006. "Formalizing Battle Management Language: A Grammar for Specifying Orders." *Paper 06S-SIW-068.* Huntsville, AL: Spring Simulation Interoperability Workshop.

SeekPNG. 2019. *SeekPNG.* Accessed 2021. https://www.seekpng.com/ipng/u2q8y3a9o0y3o0o0_however-it-wasnt-until-1997-that-machine-learning/.

Shampine, David. 2010. "The Navy's Software System Safety Technical Review Panel (SSSTRP)." *Leading Edge*, 66–67.

Soller, A., and J. Morrison. 2008. "The Effects of Automation on Battle Manager Workload and Performance." *Institute for Defense Analyses* . January. https://apps.dtic.mil/dtic/tr/fulltext/u2/a482741.pdf .

SPEC Innovations. 2021. *Innoslate Help Center.* June 12. Accessed 2021. https://help.innoslate.com/users-guide/diagrams-view/types/lml/risk-diagram/.

Tamir, Michael. 2021. "What is Machine Learning?" *Berkely School of Information.* June 26. Accessed 2021. https://ischoolonline.berkeley.edu/blog/what-is-machine-learning.

Wang, Weiyu, and Keng Siau. 2019. "Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda." *Journal of Database Management* 61–79.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California