ACADEMIA MILITAR
MILITARY ACADEMY

TÉCNICO
LISBOA

Risk Evaluation Methodology for AI-enabled Cybersecurity in Federated Mission Networks

Second Lieutenant Eduardo José da Costa Pinto Silva Esteves

Thesis to obtain the Master of Science Degree in

Military Electrical Engineering

Supervisor: Prof. Miguel Nuno Dias Alves Pupo Correia

Examination Committee

Chairperson: Prof. Pedro Nuno Mendonça dos Santos
Supervisor: Prof. José Silvestre Serra da Silva
Member of the Committee: Prof. Miguel Nuno Dias Alves Pupo Correia
Signals Lieutenant Colonel (OF-4) Pena Madeira

December 2024

"When do you think people die? When they are shot through the heart by the bullet of a pistol? No. When they are ravaged by an incurable disease? No. When they drink a soup made from a poisonous mushroom!? No! It's when... they are forgotten."

— Dr. Hiriluk One Piece

Acknowledgments

First and foremost, I would like to express my deepest gratitude to Professor Miguel Pupo Correia for his invaluable guidance, insightful advice, and continuous support throughout the course of this work. His expertise and mentorship have been instrumental in shaping this thesis, and I am profoundly grateful for his dedication and patience. Not least important, I would also like to give my thanks to Signals Major (OF-3) Luís Filipe Xavier Cavaco Mendonça Dias for his help throughout the majority of the process and coming up with the idea for this thesis.

I would also like to extend my sincere thanks to the Academia Militar, Instituto Superior Técnico and INESC-ID for providing the academic foundation and resources necessary for the completion of this thesis. To my fellow Signals course colleagues, your camaraderie and encouragement have been essential, and I feel privileged to have shared this journey with you.

Lastly, but most importantly, to my friends and family, thank you for your unwavering support, understanding, and love throughout this process. Your belief in me has been a constant source of strength and motivation

Resumo

A crescente complexidade das ameaças cibernéticas exige o desenvolvimento de ferramentas de cibersegurança mais avançadas. Esta dissertação apresenta uma metodologia de avaliação de risco para soluções de cibersegurança habilitadas por Inteligência Artificial (IA), com foco em Redes de Missão Federada (FMNs), um componente crucial nas operações militares e governamentais.

A metodologia baseia-se em metodologias consagradas, como a metodologia de classificação de risco OWASP e o modelo de ameaças STRIDE, para avaliar sistematicamente os riscos que a IA introduz nesses ambientes. O núcleo desta pesquisa foca em garantir que os sistemas de IA sejam não apenas eficazes, mas também seguros, enfatizando a importância de monitoramento contínuo, mecanismos robustos de autenticação e detecção de ameaças em tempo real em redes de missão crítica.

A metodologia é aplicada a ferramentas como OutGene e Cortex XDR, entre outras, prefazendo um total de nove ferramentas, ilustrando a necessidade de uma rigorosa avaliação de IA na proteção de infraestruturas sensíveis. Ao abordar vulnerabilidades específicas de IA e alinhar práticas de segurança com os requisitos únicos das FMNs, esta dissertação contribui para o crescente campo da cibersegurança aprimorada por IA, oferecendo uma abordagem estruturada e escalável para garantir a resiliência de redes e a integridade operacional.

Palavras-chave: Cibersegurança habilitada por IA, Redes de Missão Federada, Gestão de Riscos, Metodologia STRIDE, OWASP, Deteção de intrusões.

Abstract

The increasing complexity of cyber threats necessitates the development of more advanced cybersecurity tools. This thesis introduces an Artificial Intelligence (AI) risk evaluation methodology for AI-enabled cybersecurity solutions, focusing on Federated Mission Networks (FMNs), a crucial component in military and governmental operations. The methodology is based on established methodologies such as OWASP's risk rating methodology and the STRIDE threat model to systematically assess the risks AI introduces in these environments. The methodology evaluates AI-driven tools based on their ability to detect, mitigate, and respond to cyber threats while identifying potential vulnerabilities, such as spoofing, tampering, and denial of service.

The core of this research focuses on ensuring that AI systems are not only practical but also secure, emphasizing the importance of continuous monitoring, robust authentication mechanisms, and real-time threat detection in mission-critical networks.

The methodology is applied to tools like OutGene and Cortex XDR, amongst others, making a total of 9 tools, illustrating the need for rigorous AI evaluation in safeguarding sensitive infrastructures. By addressing AI-specific vulnerabilities and aligning security practices with the unique requirements of FMNs, this thesis contributes to the growing field of AI-enhanced cybersecurity, offering a structured, scalable approach to ensuring network resilience and operational integrity.

Keywords: AI-enabled cybersecurity, Federated Mission Networks, Risk management, STRIDE methodology, OWASP, Intrusion Detection.

Contents

List of Tables

List of Figures

Nomenclature

AI Artificial Intelligence

AI RMF AI Risk Management Framework

CIA Confidentiality, Integrity and Availability

DSR Design Science Research

DoS Denial of Service

FMN Federated Mission Network

GDPR General Data Protection Regulation

HIDS Host-based Intrusion Detection System

HIPAC Health Insurance Portability and Accountability Act

IDS Intrusion Detection System

IPDRR Identify, Protect, Detect, Respond, and Recover

ICS Industrial Control Systems

IT Information Technology

MFA Multi-Factor Authentication

ML Machine Learning

NATO North Atlantic Treaty Organization

NDR Network detection and response

NIDS Network-based Intrusion Detection System

NIST National Institute of Standards and Technology

OWASP Open Web Application Security Project

RBAC Role-Based access control

RMF Risk Management Framework

SIEM Security Information and Event Management

SOC Security Operations Center

SOAR Security Orchestration, Automation, and Response

STRIDE Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege

VPN Virtual Private Network

XDR Extended Detection and Response

Chapter 1

Introduction

This chapter will discuss the challenges and the motivation for this particular problem, define the main objectives of this work, and provide a brief thesis outline.

1.1 Motivation

With the increasing reliance on computer systems and networks in nearly every aspect of our lives, from personal banking and shopping to critical infrastructure such as power grids and transportation systems, the importance of effective cybersecurity cannot be overstated. Cyber-attacks can have far-reaching consequences, from the theft of sensitive information and financial loss to the disruption of essential services and even physical damage [1].

As a result, cybersecurity has become a crucial component of national security and is a top priority for governments, businesses, and individuals. It involves various activities, including implementing strong security policies, standards, and procedures, encrypting data, regularly patching and updating software in sensible assets, and training employees to recognize and avoid common cyber threats.

Overall, cybersecurity aims to create a secure digital environment where individuals and organizations can confidently operate, knowing their information and systems are protected against attacks.

Cybersecurity has faced several new challenges in recent years as technology advances and cyber threats become more sophisticated. Two of the key challenges have been the development of new Intrusion Detection Systems (IDS) that can effectively identify and

respond to these threats, as well as Security Information and Event Management (SIEM) solutions to help analysts from said organizations tackle and better grasp various threat scenarios.

One of the ways in which IDSs and SIEMs are evolving to meet these challenges is through the implementation of Artificial Intelligence (AI) through Machine Learning (ML) algorithms. These algorithms allow them to learn and adapt to new threats, improving their ability to identify and respond to them.

Additionally, AI can help automate many tedious and time-consuming tasks associated with managing and maintaining an IDS, freeing security personnel to focus on more strategic tasks. However, implementing AI in IDSs also presents its own set of challenges. For example, it can be difficult to ensure that the algorithms used by the IDS are free of bias, which can lead to false positives and other errors. Adding to all of this, there is a risk that attackers may be able to manipulate the AI in IDSs to bypass their defenses.

Overall, the integration of AI into cybersecurity tools has the potential to enhance their effectiveness greatly. Still, it also requires careful consideration and planning to ensure that it is implemented in a way that is both effective and secure. More importantly, AI cybersecurity tools might be attacked [2], which may also become a vulnerability for the systems they protect.

1.2 Objectives

In light of the growing reliance on AI-enabled cybersecurity tools, especially within the context of mission-critical networks like Federated Mission Networks (FMNs), a concept developed by the North Atlantic Treaty Organization (NATO), this thesis aims to explore the potential of AI tools and associated risks. As the complexity and sophistication of AI increase, it is essential to evaluate both their benefits and potential vulnerabilities when deployed in cybersecurity frameworks.

To guide this investigation, three core questions are central to this research:

- Q1: What are the critical risks of using AI in cybersecurity applications within Federated Mission Networks?
- Q2: How can we effectively evaluate AI tools in terms of their ability to secure FMNs while addressing potential vulnerabilities introduced by these technologies?

- Q3: What methodologies and frameworks can ensure AI-enabled cybersecurity tools operate securely and resiliently within high-stakes, mission-critical environments?

This research will evaluate AI's role in cybersecurity by proposing a structured risk assessment methodology, integrating well-established methodologies to assess AI's impact on security, and proposing mitigation strategies for identified risks. Through these analyses, the study seeks to create a robust methodology that ensures AI tools can be safely deployed in FMNs while addressing emerging threats and risks.

1.3 Thesis Outline

This thesis presents a risk evaluation methodology for AI-enabled cybersecurity tools, focusing on their use in FMNs. The structure of the work is designed to systematically address the complexities of cybersecurity in these environments while developing a comprehensive framework for assessing AI tools.

The introduction presents the motivation behind researching AI-enabled cybersecurity solutions in the context of FMNs. It sets the stage for the study by highlighting the increasing role of AI in addressing sophisticated cyber threats and the critical need for ensuring the security of sensitive military networks. The chapter concludes by outlining the specific objectives of the research, with a focus on developing a methodology to assess AI tools used in these environments.

Chapter 2 delves into the role of AI in cybersecurity, discussing its applications in IDS, SOAR systems, and SIEM platforms. Additionally, this chapter introduces Federated Mission Networks, exploring the security challenges they face, such as interoperability, data protection, and access control. These challenges provide the context for understanding why a rigorous evaluation of AI tools is essential in FMN environments.

Chapter 3 section reviews existing AI risk management frameworks such as OWASP, STRIDE, and NIST's AI risk management frameworks. This chapter also identifies gaps in current research, particularly in evaluating AI tools for use within FMNs, underscoring the need for a focused and robust assessment framework tailored to these specific scenarios.

Chapter 4 outlines the proposed AI risk evaluation methodology, which combines the STRIDE threat modeling approach with OWASP's risk rating methodology. The methodology aims to assess the vulnerabilities, likelihood, and impact of using AI tools in

FMNs. This chapter explains the methodology step-by-step. It discusses how it addresses the specific security concerns relevant to FMNs, making it a practical tool for cybersecurity analysis in these complex environments.

Chapter 5 applies the evaluation framework to four distinct AI tools: OutGene, Cortex XDR, Demisto (Cortex XSOAR), and Phantom (Splunk SOAR). This application demonstrates how the framework systematically identifies vulnerabilities, assesses risks, and proposes mitigation strategies for each tool in the FMN context. The chapter illustrates the practical utility of the framework through detailed case studies of these tools.

The chapter then extends the application of the methodology to additional AI-enabled cybersecurity tools, namely DarkTrace, Vectra AI, Cynet, Cybereason, and Swimlane. By comparing these tools with the ones previously studied in depth, this chapter provides a streamlined assessment based on their similarities, further showcasing the adaptability and scalability of the framework.

Finally, the conclusion summarizes the key findings of the thesis, emphasizing the importance of evaluating AI-enabled tools in Federated Mission Networks. It reflects on the efficacy of the proposed framework in identifying and mitigating risks. It suggests potential avenues for future research, such as expanding the framework to other AI technologies or civilian cybersecurity networks.

Chapter 2

Background

The following chapter will clarify some crucial topics to bear in mind during the conception of the document for a better understanding of the subject at hand.

2.1 Cybersecurity Functions (IPDRR model)

The Cybersecurity functions or the Identify, Protect, Detect, Respond, and Recover (IPDRR) model is a framework that organizations often use to guide their cybersecurity efforts:

1. Identify: Identify the assets (e.g., systems, data, networks) that need to be protected, as well as the potential risks and vulnerabilities that they are exposed to. ML algorithms could identify patterns and anomalies in network traffic and user behavior, allowing organizations to detect and track potential threats [3].

2. Protect: Implement measures to prevent or mitigate potential security incidents, such as installing firewalls, configuring access controls, and implementing security policies and procedures. Deep learning models could be used to analyze large amounts of data and learn to recognize patterns associated with malicious activity, allowing organizations to prevent attacks proactively [4].

3. Detect: continuously monitoring the organization's systems and networks for security events or incidents and responding appropriately. Natural language processing (NLP) techniques could be used to analyze text-based communication and identify indicators of compromise, such as the use of specific language or the presence of certain keywords [5].

4. Response: taking immediate action to address a security incident, such as isolating

infected systems, blocking malicious traffic, or restoring systems from backup. Decision tree algorithms could be used to automate the response process by evaluating different scenarios and determining the most appropriate course of action based on the data available [6].

5. Recovery: returning systems and data to normal operation after a security incident has been addressed and taking steps to prevent similar incidents from occurring in the future. Graph-based approaches could be used to model the relationships between different components of an organization's IT infrastructure and identify the potential impacts of an attack, allowing for more effective recovery efforts [7].

By following the IPDRR model, organizations can build a robust cybersecurity program that helps to ensure the confidentiality, integrity, and availability of their systems and data.

2.2 Artificial Intelligence and Machine Learning

AI is a 60-year-old field of Computer Science that refers to the creation of systems that can perform tasks typically requiring human intelligence, such as decision-making, problem-solving, understanding language, and recognizing patterns. AI is broadly categorized into [8]:

- **Narrow AI (Weak AI)**: AI systems designed to accomplish specific tasks, like image recognition or language translation. These systems operate under predefined parameters without general intelligence.

- **General AI (Strong AI)**: Hypothetical systems capable of understanding and learning from experiences across various tasks, mimicking human cognitive abilities.

- **Machine Learning (ML)**: A subset of AI focused on creating algorithms that enable machines to learn from data and improve over time. It includes supervised, unsupervised, and reinforcement learning techniques.

- **Deep Learning**: A subset of machine learning that uses multi-layered neural networks to analyze complex data sets. It has become particularly effective in fields like image recognition, speech processing, and autonomous systems.

AI is increasingly integrated into diverse fields, including healthcare, cybersecurity, finance, and automation, improving efficiency, reducing human errors, and enabling the automation of complex tasks.

ML is a field of AI that uses statistical techniques to allow computer systems to learn (i.e., progressively improve performance on a specific task) from data without being explicitly programmed. ML algorithms build a mathematical model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed. ML aims to create machines that can automatically learn how to make decisions. This learning process involves training a computing device to analyze a set of data (called training data) using a specific ML algorithm. [9].

There are several types of ML. Three that have been applied in the context of cybersecurity [10] are supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning uses labeled data to teach a machine how to perform a task. Labeled data is data that has already been classified into one or more categories. Examples of supervised learning include image recognition, facial recognition, and natural language processing.

- Unsupervised learning uses unlabeled data to allow a machine to identify patterns in data. Unsupervised learning does not require labeled data and is used for clustering, dimensional reduction, and anomaly detection tasks.

- Reinforcement learning is a type of machine learning that uses reward and punishment to teach a machine how to perform a task. Reinforcement learning algorithms use trial and error to learn from their mistakes and adjust their behavior accordingly. Examples of reinforcement learning include game-playing, robotics, and autonomous vehicles.

With the help of ML tools, it is possible to enhance certain capabilities of cybersecurity functions, making it the target of recent studies around the matter [11].

It is essential to highlight how ML techniques significantly enhance the effectiveness and scope of cybersecurity operations. ML perfectly complements cybersecurity functions with its capacity to process large volumes of data, recognize patterns, and adapt over time. ML fundamentally transforms these functions by automating tasks that traditionally require manual intervention, offering higher precision and speed in threat detection,

mitigation, and response.

For instance, supervised learning models can be used to enhance the Identify function by recognizing complex patterns in network traffic that may indicate potential vulnerabilities or cyber-attacks. This proactive approach allows for the early detection of security threats. On the other hand, unsupervised learning algorithms, such as clustering techniques, significantly contribute to the Detect function by identifying anomalies that deviate from normal network behavior and flagging potential threats that may not have been previously encountered or classified.

Moreover, ML also strengthens the Respond and Recover functions by automating threat response workflows. For example, decision tree models can be applied to guide automated systems in taking immediate and appropriate actions during active threats. In parallel, graph-based approaches help map relationships between network components, aiding faster and more efficient incident recovery.

Therefore, integrating ML into cybersecurity functions creates a robust, scalable solution for addressing the increasing complexity of cyber threats. This seamless blend of ML and cybersecurity not only enhances the ability to predict and identify risks but also enables both preventive and reactive measures across the entire cybersecurity landscape.

2.3 Security Systems

This section presents a sample of security services or tools that can rely on AI.

2.3.1 Intrusion Detection Systems

An IDS is a security system that monitors a network or system for malicious activity or policy violations. It detects malicious activity and generates an alert so that appropriate actions can be taken. There are two main types of IDSs [12]: network-based IDSs and host-based IDSs:

- Network-based IDSs (NIDS) are designed to monitor network traffic for signs of security threats. They typically operate at the network layer and can analyze traffic from multiple hosts at once. NIDSs are useful for detecting threats that are spread across a network, such as malware or denial of service attacks.

- Host-based IDSs (HIDS) are designed to monitor a single host for signs of security threats. They operate at the host level and are useful for detecting threats that are specific to a single host, such as unauthorized access or changes to system files.

There are also two approaches, among others, to intrusion detection: Anomaly-based Intrusion Detection and Signature-based Intrusion Detection. Anomaly-based detection uses heuristic rules or statistical models to classify behavior as benign or malicious, while signature-based detection looks for predefined patterns to identify known attacks.

Oliveira et al. [12] discuss using artificial intelligence algorithms, such as Neural Networks, Random Forests, k-Nearest Neighbors, and Support Vector Machines, to improve intrusion detection performance. There were also some experiments using three different artificial intelligence models (Random Forests, Multilayer Perceptron, and Long Short-Term Memory) on two datasets (CIDDS-001 and UNSW-NB15) to evaluate the performance of different anomaly detection approaches. It was found that the multi-flow approach and the Long Short-Term Memory model achieved the best performance in terms of both detection and attack classification, turning itself into a great example of AI applied to cybersecurity solutions.

2.3.2 Security Information and Event Management

SIEMs are tools used by SOCs to collect, normalize, and analyze security events from various sources [13]. These events may be generated by Information technology (IT) assets such as networks, perimeter defense systems, and application servers. They can include log files and activity from sensors monitoring these assets. When a SIEM system detects a possible malicious activity, it triggers an alert, which is then reviewed by SOC personnel to determine if further action, such as coordinating incident response and forensic activities, is necessary. SIEMs use rule engines to apply rules to events and generate alerts, which can be created by analysts or algorithmically generated from events through pattern mining or anomaly detection techniques. The effectiveness of a SIEM system depends on its ability to ingest and analyze large amounts of data, as well as its access to actionable threat intelligence

Despite being a strong technical foundation for SOCs, SIEMs still have room for improvement in adaptability, context awareness, flexibility, holism, and social integration.

Several potential AI-based approaches could address these issues in SIEM systems. For

example, ML techniques could improve the adaptability and context-awareness of SIEMs by enabling them to learn from past events and adapt their behavior accordingly[14]. Natural language processing could improve the social integration of SIEMs by enabling them to understand and analyze text-based communications and extract relevant information [5]. Finally, graph-based approaches could improve the holism and flexibility of SIEMs by allowing them to model complex relationships and connections between events and entities [7].

2.3.3 Security Orchestration Automation Response

Security Orchestration, Automation and Response (SOAR) [15] is a set of tools and technologies organizations use to automate and improve their cybersecurity operations. SOAR helps organizations manage and respond to cybersecurity threats more efficiently and effectively by automating the collection and analysis of security data, the creation and execution of response plans, and the communication of relevant information to stakeholders.

SOAR technologies often include a security orchestration platform, which coordinates the actions of different cybersecurity tools and systems, and an automation engine, which automates the execution of tasks and processes. SOAR solutions may also include a security incident response platform, which helps organizations track and manage the progress of their response to a cybersecurity threat.

Some of the most popular examples of SOAR platforms include Demisto, Phantom, and Swimlane, which provide tools for security analysts to collect and analyze data from various sources and use that information to identify and respond to potential threats. It also includes functionality for automating certain incident response tasks. They mainly provide the following range of features:

- Automation of incident response tasks
- Integration with a wide range of security technologies and platforms
- Collaboration and communication tools for security teams
- A library of security playbooks and use cases
- A machine learning-based threat intelligence engine

2.4 Federated Mission Networks

A segregated network is a type of computer network that is physically or logically isolated from other networks, both internally and externally. This means the segregated network is disconnected from other networks and operates independently. Segregated networks are typically used in environments where security is paramount, such as in the military, government, and financial organizations, where sensitive information must be protected from unauthorized access and cyber-attacks.

A segregated network may also be called an air-gapped, isolated, or closed network [16]. The network may be physically segregated by using separate hardware or software components or logically segregated using access controls, firewalls, or other security measures to prevent unauthorized access. Sometimes, a segregated network may be connected to other networks but only through a limited number of controlled access points, such as a secure gateway or Virtual Private Network (VPN).

A FMN is a concept that refers to a collaborative and interoperable network infrastructure used by different organizations, agencies, or entities to support joint operations, missions, or activities. The primary goal of an FMN is to enable secure and seamless information sharing and communication among diverse participants while maintaining data integrity and confidentiality [17].

The FMN approach recognizes that modern operations often involve multiple entities, such as military forces, government agencies, allied nations, and partner organizations. These entities may have independent networks, systems, and information repositories. However, to effectively collaborate and achieve shared objectives, it is crucial to establish a common network infrastructure that allows authorized participants to connect, communicate, and exchange information in a controlled and coordinated manner.

In a FMN, participating organizations retain control over their network resources and assets while adhering to common standards and protocols. This enables secure and seamless integration of disparate systems, applications, and data sources, regardless of the specific technologies each participant employs. The FMN promotes information sharing, situational awareness, and decision-making across organizational boundaries, facilitating better coordination, efficiency, and effectiveness in joint missions or operations.

The key features of a FMN typically include:

- Interoperability: The FMN ensures that diverse systems and networks can commu-

nicate and exchange information effectively by adopting standardized protocols and interfaces.

- Security: Robust security measures, including authentication, encryption, access control, and data protection, are implemented to safeguard sensitive information and prevent unauthorized access or data breaches.

- Scalability: FMNs are designed to accommodate a large number of participants and handle increasing volumes of data, ensuring the network can scale as mission requirements evolve.

- Flexibility: The FMN architecture allows participants to connect using a variety of technologies, such as wired, wireless, or satellite communication, accommodating different operational environments and scenarios.

- Governance: FMNs require clear governance structures, policies, and procedures to establish rules for participation, information sharing, data ownership, and operational guidelines.

Overall, a FMN serves as a foundation for enhanced collaboration, coordination, and information sharing among diverse entities involved in joint operations, enabling them to work together more effectively towards common goals while preserving the autonomy and security of each participant's network infrastructure.

Within the FMN concept, there is a term called Spirals. It refers to incremental development and deployment phases within the FMN framework [18]. Each spiral represents a specific phase or iteration of capability development and deployment. Some key characteristics and differences between the spirals are the following:

- Spiral 1:
  - Early phase of FMN development.
  - Focuses on establishing foundational infrastructure, network connectivity, and basic interoperability.
  - Aims to establish a common operating environment and enable basic information sharing between participating entities.

- Spiral 2:
  - Builds upon the achievements of Spiral 1.
  - Focuses on enhancing information sharing and collaboration capabilities.
  - Expands interoperability and introduces more advanced communication and data exchange mechanisms.
  - Incorporates security enhancements and policies to protect sensitive information.
- Spiral 3:
  - Further advances information sharing and collaboration capabilities.
  - Emphasizes integration of additional mission-specific applications and services.
  - Enhances security measures, including identity management and access controls.
  - Expands the scope of FMN to involve a broader range of mission partners and stakeholders.
- Spiral 4:
  - Represents a significant milestone in FMN maturity.
  - Focuses on achieving a highly interoperable and scalable network infrastructure.
  - Enables seamless integration and interoperability across multiple domains and mission areas.
  - Incorporates advanced technologies and standards to support complex mission requirements.

It's important to note that each spiral's specific characteristics and scope may vary depending on the context and specific implementation of FMN. The spirals provide a structured and incremental approach to capability development, allowing for continuous improvement and expansion of FMN capabilities over time.

Following NATO's roadmap of implementing these spirals, we are currently in the spiral 4 preferred operational use phase. This information will help shape the guidelines and procedures while developing our methodology.

As of today, other spirals are being developed and implemented, but due to the nature of the documentation, we kept them out of the scope of this thesis.

Chapter 3

Related Work

In this chapter, we will explore the various ways AI is being used in cybersecurity in the current literature and risk management methodologies currently employed in organizations.

3.1 Artificial Intelligence in Cybersecurity

AI has the potential to revolutionize cybersecurity by enabling computers to detect and respond to cyber threats in real-time. AI algorithms can analyze vast amounts of data and recognize patterns humans might miss, making them well-suited for identifying and mitigating cyber attacks. There are several ways in which AI is being used in cybersecurity, including:

- Intrusion detection: AI algorithms can analyze network traffic and identify anomalies that may indicate a cyber attack.
- Vulnerability assessment: AI algorithms can analyze software and identify vulnerabilities that cyber attackers could exploit.
- Malware detection: AI algorithms can analyze files and identify malware that may be hidden within them.
- Phishing detection: AI algorithms can analyze emails and identify phishing attacks that may trick users into divulging sensitive information.

In Apruzzese et al. [19], a comprehensive overview and practical recommendations for the use of ML in cybersecurity are given, with a focus on making the information

accessible to a wide range of stakeholders, including security specialists, executives, and researchers. It also aims to clear misconceptions about the use of ML in cybersecurity, and it distinguishes itself from previous work by offering a "meta-review" that provides a comprehensive overview, practical recommendations, and research directions and addresses misconceptions within the entire cybersecurity sphere.

Broad examples are given in terms of ML's application in the cases of Threat detection, whether Network Intrusion, Malware, or Phishing, showing promising results, Alert Management, Raw data Analysis, Risk Exposure Assessment, and others. All of these show an excellent omen towards the use of AI in cybersecurity, but it also comes with its disadvantages as the cybersecurity ecosystem is highly dynamic, and it becomes challenging to provide real-world data to these AI solutions to be trained as well as companies tending to not sharing their own data making this an even harder spot for AI to thrive.

Some work has to be done to create transparency and disclosure of said environment data to compete with the shortage of adequate data, making AI tools much closer to the real world as possible to better help SOC teams perform their jobs and keep organizations, critical or non-critical, safe.

3.2 Risk Management

Risk assessment and management play crucial roles in cybersecurity and beyond. They are systematic processes designed to identify, evaluate, and mitigate potential risks organizations face in achieving their objectives. By understanding and effectively managing risks, organizations can make informed decisions, allocate resources wisely, and protect their assets and interests.

Kaplan and Garrick [20] explored the concept of risk and provided insights into its quantitative definition. Risk, in the context of this paper, refers to the probability of an undesirable event or outcome occurring and its potential impact on objectives or interests:

- Likelihood: Risk involves assessing an event or outcome's likelihood. This probability can range from low to high, indicating the chance of the event happening. Probability is often expressed numerically or as a percentage, representing the frequency or likelihood of occurrence.

- Impact: Risk also considers the potential consequences or impact that may result

from the event's occurrence. The impact can vary in severity, from minor disruptions to significant loss or harm. It encompasses various dimensions such as financial, operational, reputational, or health and safety impacts.

The main objective here is never 100% security but acceptable risk, which can translate into defining risk as a relation likelihood of a security violation and the impact of said security violation:

$$risk = likelihood \times impact \tag{3.1}$$

The likelihood of a security violation is the relation between the threat level and the vulnerability level:

$$likelihood = threat \times vulnerability \tag{3.2}$$

Most risk rating methodologies are based on these two principles, which is why we will look at some examples in this section.

3.2.1 OWASP Risk Rating Methodology

The Open Web Application Security Project (OWASP) developed the OWASP Risk Rating Methodology that offers a structured approach to risk assessment, enabling cybersecurity analysts to manage and mitigate potential threats effectively.

The OWASP Risk Rating Methodology consists of six main steps that guide analysts through the risk assessment [21]:

- Step 1: Identifying a Risk: The first step involves identifying and documenting the risks associated with software vulnerabilities. This includes considering various factors such as the nature of the vulnerability, the potential impact it may have, and the likelihood of exploitation.

- Step 2: Factors for Estimating Likelihood: In this step, analysts evaluate the likelihood of a vulnerability being exploited. Factors such as the prevalence of the vulnerability, the skill level of potential attackers, and the effectiveness of existing security controls are taken into account to estimate the likelihood accurately.

- Step 3: Factors for Estimating Impact: The next step focuses on assessing the potential impact of a successful exploitation. Analysts consider the potential harm

to the confidentiality, integrity, and availability of the system or data, as well as any associated business, regulatory, or reputational impacts.

- Step 4: Determining Severity of the Risk: Based on the estimated likelihood and impact, analysts determine the severity of each identified risk. This step helps prioritize the risks by assigning them appropriate severity levels.

- Step 5: Deciding What to Fix: Once the risks are categorized by severity, analysts can make informed decisions on which risks to address first. They consider the severity level, available resources, and the organization's overall risk tolerance.

- Step 6: Customizing The Risk Rating Model: The final step involves customizing the risk rating model to suit the organization's specific needs. This may include adjusting the factors used to estimate likelihood and impact or incorporating additional considerations based on the organization's risk appetite.

By following these six steps, the OWASP Risk Rating Methodology enables cybersecurity analysts to systematically assess and prioritize software vulnerabilities. It empowers organizations to allocate resources effectively, focus on critical risks, and implement targeted risk mitigation strategies.

3.2.2 Windows STRIDE and DREAD Threat Modelling

STRIDE and DREAD are two models mainly used in threat modeling that provide structured approaches for identifying and evaluating potential threats to software systems [22].

STRIDE is an acronym representing a set of threat categories used in threat modeling. It stands for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. STRIDE provides a systematic approach to identifying and evaluating potential threats during the software development lifecycle.

Each letter in STRIDE corresponds to a specific type of threat:

- Spoofing: This threat category involves an attacker impersonating another user or entity to gain unauthorized access or deceive the system.

- Tampering: Tampering threats involve unauthorized modification, alteration, or destruction of data or system components, leading to potential security breaches.

- Repudiation: Repudiation threats relate to the inability to verify or authenticate actions or events, making it difficult to prove or disprove the involvement of specific individuals or entities.

- Information Disclosure: Information disclosure threats refer to unauthorized access or exposure of sensitive information, potentially leading to privacy breaches or data misuse.

- Denial of Service: Denial of Service threats aim to disrupt or degrade the availability or functionality of a system, making it inaccessible or unusable for legitimate users.

- Elevation of Privilege: Elevation of Privilege threats involve unauthorized escalation of user privileges, allowing attackers to gain higher levels of access or control within a system.

By systematically analyzing a system or software application using the STRIDE framework, security professionals can identify potential vulnerabilities and design appropriate countermeasures to mitigate the identified threats.

DREAD is another method that focuses on assessing and prioritizing risks based on five key factors: Damage potential, Reproducibility, Exploitability, Affected users, and Discoverability. It provides a structured approach for evaluating the severity and impact of identified threats.

The factors in the DREAD framework are as follows:

- Damage potential: This factor assesses a threat's potential impact and severity if it were to be successfully exploited. It considers the potential harm to assets, data, and the overall system.

- Reproducibility: Reproducibility evaluates how easily an attacker can replicate or exploit the identified threat. It takes into account the effort required to exploit the vulnerability consistently.

- Exploitability: Exploitability measures the ease with which an attacker can exploit the vulnerability associated with the threat. It considers factors such as the availability of tools, knowledge, and skills required to carry out the attack.

- Affected users: This factor assesses the number of users or entities impacted by the threat. It considers the potential reach and scope of the vulnerability in terms of affected individuals or systems.

- Discoverability: Discoverability evaluates the likelihood of the threat being discovered by potential attackers or security researchers. It considers factors such as the visibility of the vulnerability and the level of attention it may attract.

By assigning scores or ratings to each factor, DREAD helps prioritize the identified threats based on their severity and potential impact. This allows organizations to focus their resources and efforts on addressing the most critical threats first.

Both STRIDE and DREAD methods provide structured approaches to identify, assess, and prioritize threats during the software development process. By systematically evaluating potential threats, organizations can proactively design and implement appropriate security measures to mitigate risks and enhance the overall security posture of their systems and applications.

3.2.3 NIST Special Publication 800-37 Revision 2

NIST Special Publication 800-37 Revision 2 (NIST.SP.800-37r2) provides guidance for managing risks associated with the security and privacy of information systems [23]. This Risk Management Framework (RMF) emphasizes integrating security and privacy considerations throughout the system development life cycle.

The main aspects of NIST.SP.800-37r2 can be summarized in four major topics:

- **Risk-Based Approach**: The RMF promotes managing risks holistically, focusing on the organization's operations, assets, and individuals rather than relying solely on technical controls.
- **Integration of Security and Privacy**: It incorporates security and privacy into the RMF, recognizing that both are crucial in modern systems.
- **Continuous Monitoring**: A major focus is the continuous assessment and monitoring of controls to ensure they remain effective over time.
- **Alignment with NIST Cybersecurity Framework (CSF)**: The RMF is aligned with the NIST CSF [24] to ensure complementary usage and help organizations meet regulatory requirements, either national or international.

Going into detail, the RMF is a comprehensive, seven-step process designed to integrate security and privacy throughout the lifecycle of information systems. It starts with preparing and categorizing systems based on risk impact, then selecting, implementing, and assessing security controls before authorizing system operation. The final step emphasizes continuous monitoring to ensure controls remain effective over time. This framework addresses an organization's entire risk management approach, blending compliance, system operations, and ongoing risk mitigation:

1. **Prepare**: Organizational preparation for risk management.
2. **Categorize**: Classify information systems and data based on the impact levels of confidentiality, integrity, and availability (CIA).
3. **Select**: Choose appropriate security and privacy controls from NIST SP 800-53 [25].
4. **Implement**: Deploy and configure the selected controls.
5. **Assess**: Review the effectiveness of the controls.
6. **Authorize**: A senior official makes a risk-based decision to authorize system operation.

7. **Monitor**: Continuously monitor the system for emerging risks and changes.

While RMF offers a broad lifecycle approach for managing risk across an organization's systems, OWASP focuses on identifying and addressing specific vulnerabilities within web applications. Both approaches are vital in their respective areas—NIST for full-scale system security and compliance and OWASP for mitigating practical software and web development risks.

3.2.4 NIST AI Risk Management Framework

The NIST AI Risk Management Framework (AI RMF) 1.0. acknowledges the significant potential of AI technologies to impact society positively. It also recognizes the risks they pose and aims to help organizations manage these risks and promote responsible development and use of AI systems [26].

The AI RMF is designed to provide organizations with approaches to enhance trustworthiness and foster responsible AI design, development, deployment, and use. It is intended to be flexible, voluntary, and applicable across sectors and use cases.

The Framework consists of two parts: framing AI risks and trustworthiness and the "Core" functions - GOVERN, MAP, MEASURE, and MANAGE - which help organizations address AI risks practically.

This framework is an upgrade from the notions introduced with OWASP methodology and STRIDE and DREAD due to its incorporation of the concepts discussed previously, but also it sees the whole risk management through a wider scope, analyzing not only the tools used and the risk to the organization but also the impacts of the environment of said organization.

The aspect of framing AI risks and trustworthiness by discussing the characteristics of trustworthy AI systems and providing guidance on how to address them. These characteristics include being valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.

Trustworthy AI systems need to be responsive to various criteria that are valuable to interested parties. Enhancing AI trustworthiness can help reduce negative AI risks. However, addressing each characteristic individually does not guarantee trustworthiness, as

tradeoffs may be involved, and the importance of each characteristic can vary in different situations.

The characteristic of being valid and reliable refers to the accuracy and robustness of AI systems. Validation confirms that the requirements for a specific use have been fulfilled, while reliability relates to the system's ability to perform as required without failure. Measures of accuracy, robustness, and reliability contribute to the trustworthiness of AI systems.

The AI RMF Core functions, as outlined in the document, consist of four main functions: GOVERN, MAP, MEASURE, and MANAGE. These functions provide a framework for managing AI risks and developing trustworthy AI systems. Here's an overview of each function and its categories/subcategories:

GOVERN Function: The GOVERN function of the AI RMF Core cultivates a risk management culture, outlines processes and documents, assesses potential impacts, aligns with organizational principles, connects technical aspects with values, and addresses the full product lifecycle. It establishes a foundation for effective AI risk management and ensures that AI systems are developed and deployed in line with organizational priorities:

- GOVERN 1: Policies, processes, procedures, and practices related to mapping, measuring, and managing AI risks.
- GOVERN 2: Accountability structures to empower and train teams and individuals responsible for managing AI risks.
- GOVERN 3: Workforce diversity, equity, inclusion, and accessibility processes.
- GOVERN 4: Organizational commitment to a risk-focused culture.
- GOVERN 5: Robust engagement with relevant AI actors.
- GOVERN 6: Policies and procedures to address AI risks arising from third-party software, data, and supply chain issues.

MAP Function: The MAP function in the AI RMF Core establishes the context for AI risks. It identifies and understands the intended purposes, potential impacts, and settings of AI systems. It incorporates interdisciplinary perspectives, considers organizational goals, and assesses the risks and benefits associated with AI technology. The

MAP function provides crucial knowledge to inform decision-making throughout the AI lifecycle:

- MAP 1: Establishing and understanding the context.
- MAP 2: Categorization of the AI system.
- MAP 3: Understanding AI capabilities, goals, benefits, and costs.
- MAP 4: Mapping risks and benefits of AI system components.
- MAP 5: Characterizing impacts on individuals, groups, communities, organizations, and society.

MEASURE Function: The MEASURE function in the AI RMF Core employs various methods and metrics to analyze, assess, and measure AI risks. It involves quantitative, qualitative, or mixed-method approaches to evaluate the functionality, trustworthiness, and performance of AI systems. Through rigorous testing, verification, and validation processes, organizations gather objective data to determine the system's safety, reliability, transparency, privacy, fairness, and environmental impact. The MEASURE function provides valuable insights for managing AI risks and informs decision-making throughout the AI system's lifecycle.:

- MEASURE 1: Identifying appropriate methods and metrics.
- MEASURE 2: Evaluating AI systems for trustworthy characteristics.
- MEASURE 3: Tracking identified AI risks over time.
- MEASURE 4: Gathering and assessing feedback about measurement efficacy.

MANAGE Function: The MANAGE function allocates resources and responds to AI risks. It involves prioritizing and managing identified risks based on their impact and likelihood. Organizations develop strategies to maximize AI benefits while minimizing negative impacts. Risk treatment options, such as mitigation, transfer, avoidance, or acceptance, are employed. The MANAGE function also includes planning for incident response, recovery, and communication. It ensures that AI systems are effectively managed throughout their lifecycle.:

- MANAGE 1: Prioritizing, responding to, and managing AI risks.
- MANAGE 2: Planning, preparing, implementing, and documenting strategies to maximize benefits and minimize negative impacts.
- MANAGE 3: Managing AI risks and benefits from third-party entities.
- MANAGE 4: Documenting and monitoring risk treatments, response and recovery plans, and communication plans.

These functions provide a structured approach to AI risk management and can be applied iteratively throughout the AI system lifecycle. Organizations can choose the relevant categories and subcategories based on their needs and resources while ensuring compliance with legal, regulatory, and ethical requirements.

3.2.5 ISO 31000: Risk Management Framework

ISO 31000 is a family of standards that provides a structured approach to managing risks across organizations of any size or sector. It emphasizes integrating risk management into all processes and decision-making [27]. The framework is guided by principles such as ensuring that risk management adds value, is customized to the organization's context, and fosters continuous improvement.

The risk management process under ISO 31000 includes identifying risks, assessing their likelihood and impact, and developing strategies to manage them, such as mitigation or acceptance. Continuous monitoring and regular reviews are crucial to adapting to changing environments. Strong leadership and clear communication with stakeholders are also essential for effective implementation.

Overall, ISO 31000 promotes a comprehensive and flexible risk management approach, helping organizations address uncertainties and enhance decision-making.

3.3 AI-Enabled Cybersecurity Tools

AI-enabled cybersecurity tools are typically trained on large datasets of network traffic and security incidents to identify patterns and anomalies that may indicate a security threat. The training process can be performed using supervised or unsupervised machine

learning algorithms, depending on the type and amount of data available. To adapt the AI model to a specific network, the cybersecurity tool may be configured with network-specific parameters, such as IP addresses, subnets, and protocols, that are relevant to the network being protected. This allows the tool to focus on the specific network traffic patterns and behaviors that are most relevant to the organization.

AI-enabled cybersecurity tools may also include active mechanisms to detect and prevent attacks on the AI model itself. One common approach is to use adversarial training, where the model is trained on both clean and adversarial examples to make it more robust to attacks. Another approach is to use anomaly detection techniques to identify unexpected behavior in the model output, which may indicate that the model has been compromised or tampered with. In addition, some AI-enabled cybersecurity tools may use explainability or transparency techniques to help detect attacks on the model. These techniques allow the tool to provide insights into how the model makes decisions, which can help cybersecurity analysts identify and mitigate potential attacks on the model.

There are several AI-enabled cybersecurity tools that are designed especially for segregated networks. Some of the most popular are:

- Darktrace: Darktrace is an AI-powered cybersecurity platform designed to detect and respond to advanced cyber threats in segregated networks [28]. The platform uses machine learning algorithms to analyze network traffic and identify anomalous behavior that may indicate a security breach.

- Vectra AI: Vectra AI is a network detection and response (NDR) platform that uses AI to detect and respond to cyber threats in segregated networks [29]. The platform is designed to monitor network traffic in real-time and identify potential threats using machine learning algorithms.

- Cortex XDR: Palo Alto Networks offers a range of AI-enabled cybersecurity tools, including its Cortex Extended Detection and Response (XDR) platform [30]. Cortex XDR is designed to provide advanced threat detection and response capabilities for segregated networks, using machine learning and behavioral analytics to identify and respond to threats.

- Cynet: Cynet is an AI-powered cybersecurity platform that provides advanced threat detection and response capabilities for segregated networks [31]. The platform

uses machine learning algorithms to analyze network traffic and identify potential threats, and it also includes automated response capabilities to help contain and remediate security incidents.

- Cybereason: Cybereason is an AI-powered endpoint protection platform that is designed to provide advanced threat detection and response capabilities for segregated networks [32]. The platform uses machine learning algorithms to analyze endpoint data and identify potential threats, and it also includes automated response capabilities to help contain and remediate security incidents.

- OutGene: OutGene is an AI-enabled cybersecurity tool that uses machine learning to analyze malware genetic code and detect and neutralize threats [33]. It adapts over time and aims to provide faster and more accurate malware detection. It leverages machine learning techniques, specifically clustering algorithms and genetic algorithms, to detect anomalies and potential cyber-attacks in network traffic data while using AI to automatically extract relevant features from the data, perform clustering analysis, and identify outliers or suspicious entities.

- Demisto: Demisto, currently known as Cortex-XSOAR, is a SOAR platform developed by Palo Alto Networks [34]. It leverages artificial intelligence to enable intelligent automation and collaboration among security teams. Demisto uses AI capabilities to automate repetitive tasks, analyze security alerts, and provide incident response playbooks, allowing organizations to streamline their security operations and respond more effectively to threats.

- Phantom: Phantom, a SOAR platform from Splunk, harnesses the power of AI to automate and standardize security workflows [35]. It integrates with existing security technologies, collects and analyzes security data, and executes automated actions based on predefined playbooks. Phantom's AI capabilities enable organizations to detect and respond to security incidents faster, reduce manual effort, and improve overall incident response efficiency.

- Swimlane: Swimlane is a SOAR platform incorporating AI-driven automation and machine learning [36]. It integrates with various security tools and data sources, analyzes security alerts and events, and automates response actions based on prede-

fined workflows. Swimlane's AI features help organizations enhance their incident response capabilities, accelerate decision-making, and proactively address security threats.

In table 3.1, we summarized each tool's advantages and disadvantages and, if they apply any, mechanisms implemented to mitigate attacks against the tools model.

Tool	**Pros**	**Cons**	**Self-Defense Mechanism**
OutGene	Uses AI to analyze malware genetic code	Limited information available	Not specified
DarkTrace	Provides real-time threat detection	Requires fine-tuning for optimal performance	Self-defends by analyzing network anomalies and blocking suspicious activity
Vectra AI	Offers advanced threat hunting capabilities	Can be complex to implement and manage	Behavioral analysis and threat intelligence to identify and block threats
Cortex XDR	Comprehensive network security platform	High cost and complex deployment	Intrusion prevention systems, firewall rules, secure management protocols
Cynet	Offers autonomous breach protection	Limited scalability for large environments	Intrusion detection, prevention systems, and automated response
Cybereason	Provides behavior-based threat detection	Initial setup and tuning can be time-consuming	Behavior-based detection, continuous monitoring, and active threat hunting
Demisto	Provides automation and orchestration of security tasks	Initial learning curve for setup and customization	Access controls, authentication mechanisms, and monitoring for suspicious activity
Phantom	Offers playbooks for automated incident response	Requires integration with other security tools	Access restrictions, encryption, and monitoring of infrastructure
Swimlane	Provides automation and orchestration capabilities	Requires customization for specific use cases	Continuous monitoring and threat hunting to respond to attacks

Table 3.1: Comparison of AI Tools: Pros, Cons, and Self-Defense Mechanisms

Chapter 4

AI Risk Evaluation Methodology

This chapter outlines the methodology that we propose to analyze, assess, and categorize the risks associated with using AI-enabled cybersecurity tools in the FMN environment. To create a robust methodology, we must first address the core aspects of security that are critical in FMN, ensuring that our methodology aligns with these networks' stringent requirements and security policies. Given the sensitive nature of FMN, adopting AI introduces a new range of security challenges, and it is essential to thoroughly evaluate these tools to safeguard the integrity and confidentiality of mission-critical data.

The methodology presented here draws from well-established risk assessment methods, such as STRIDE and the OWASP Risk Methods, recognized for their effectiveness in identifying, evaluating, and mitigating cybersecurity threats. By leveraging these methods, we ensure that our risk evaluation methodology is comprehensive and adaptable, providing a reliable tool for cybersecurity professionals working within FMN environments. This chapter will explain the rationale behind selecting these models, their application in our proposed methodology, and the steps involved in implementing them. We aim to offer a structured and scalable approach that can be consistently applied across different AI tools and FMN scenarios, making it the go-to methodology for evaluating AI in this context.

After outlining the fundamental requirements of FMN security, we will discuss how our risk evaluation methodology incorporates aspects of the OWASP and STRIDE methods. These approaches have been tested and proven in various cybersecurity contexts. By integrating them into our methodology, we aim to create a practical and effective tool for assessing the risks posed by AI in FMN environments.

4.1 AI Risk Evaluation Methodology Requirements

Implementing FMNs presents a unique challenge, as it involves connecting multiple, often disparate, networks and systems to enable seamless communication and information sharing across mission-critical environments. To ensure the security of these networks, the risk evaluation methodology must address several core requirements essential for the successful operation of FMNs.

First, it is critical to clearly define the mission objectives and the specific requirements for FMN usage. These include adhering to cybersecurity policies, standards, and best practices established by organizations such as NATO and NIST. Security controls must be implemented to ensure that data integrity, confidentiality, and availability are always maintained. Furthermore, by adhering to defined technical and security standards, the methodology must ensure seamless and secure interoperability between various systems, including those of different nations and organizations.

Robust authentication and authorization mechanisms must also be established to prevent unauthorized access. Role-based access control (RBAC) and multi-factor authentication (MFA) are security mechanisms that could enhance the overall security posture of FMNs. In addition, continuous monitoring and real-time threat detection are essential to maintaining the integrity and security of FMN environments. These requirements serve as the foundation for our AI evaluation methodology, guiding the selection of criteria and methodologies we will apply.

In the context of FMNs, decisions made regarding security must always align with these requirements to ensure that the network remains resilient against emerging threats. By incorporating these principles into our methodology, we provide a robust foundation for evaluating the risks posed by AI tools and mitigating vulnerabilities effectively in mission-critical environments. Ultimately, the success of FMN relies on the ability to securely share information and communicate across networks without compromising the mission's objectives or the safety of its participants.

4.1.1 Design Science Research

The AI risk evaluation methodology was developed using a rigorous methodology, which is Design Science Research (DSR). DSR is a research paradigm focused on the creation

and evaluation of artifacts—whether models, systems, methods, or constructs—that are designed to solve identified problems. Unlike traditional research methodologies, which often focus on describing phenomena or testing hypotheses, DSR emphasizes the development of practical, innovative solutions. These artifacts address real-world problems and contribute to advancing research and knowledge in the field.

According to Peffers et al.[37], DSR follows a structured six-stage process:

1. **Problem Identification and Motivation**: Define the research problem clearly and justify why solving this problem is essential. This step provides the foundation for developing the artifact.

2. **Objectives of a Solution**: Establish clear goals for what the solution must achieve based on the problem context. These objectives will guide the design of the artifact.

3. **Design and Development**: Create the artifact, a model, method, or system to address the identified problem. This stage involves both theoretical design and practical implementation.

4. **Demonstration**: Test the artifact by using it to solve the identified problem. This can involve simulations, case studies, or experiments demonstrating the artifact's efficacy.

5. **Evaluation**: Measure the artifact's performance by comparing it to the established objectives. Evaluation criteria may include usability, effectiveness, and scalability.

6. **Communication**: Disseminate the research findings to academic and professional communities, ensuring that the knowledge gained from developing the artifact is accessible and can be applied in practice.

In this thesis, DSR is an appropriate framework because the goal is to develop an AI risk evaluation methodology that not only addresses theoretical concerns but also provides practical, actionable solutions for organizations integrating AI into their cybersecurity operations. The artifact, in this case, is the risk evaluation methodology itself, which will undergo rigorous design, development, and testing to ensure that it meets the unique security challenges presented by AI in FMN environments. Following the DSR process ensures the methodology is scientifically sound and readily applicable in real-world scenarios.

The iterative nature of DSR further aligns with the objectives of this thesis, as it allows for continuous refinement of the methodology based on the results of demonstration and evaluation. This ensures that the final artifact—the AI risk evaluation methodology—is practical and adaptable to new challenges as AI technologies evolve.

4.2 AI Risk Evaluation Methodology Outline

In response to the multifaceted security challenges outlined in Chapter 3, developing a comprehensive AI risk evaluation methodology that leverages proven methodologies is essential. Our AI risk evaluation methodology combines two of the most respected models in cybersecurity: the STRIDE Threat Model and the OWASP Risk Method. By integrating these two approaches, we provide a structured and systematic method for assessing and mitigating the security risks associated with AI-enabled tools within the FMN environment.

The STRIDE method serves as the cornerstone for identifying and categorizing potential vulnerabilities introduced by AI tools. Originally developed by Microsoft, STRIDE offers a well-established method that allows for a detailed analysis of the security threats AI systems may pose across six critical categories: Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. In the context of FMNs, where secure and efficient communication between multinational entities is paramount, this level of granularity is crucial. By systematically mapping AI tool functionalities to each STRIDE category, we can anticipate a broad spectrum of threats that may compromise the integrity, confidentiality, and availability of mission-critical systems. STRIDE not only identifies the types of threats but also provides a structured way to understand how these threats could manifest, enabling us to evaluate how AI tools might be exploited in this highly sensitive environment.

While STRIDE provides a comprehensive approach to threat identification, the OWASP Risk Methodology complements it by introducing a structured process for assessing and prioritizing the risks associated with each identified threat. The OWASP method guides us through a methodical assessment of the likelihood of exploitation and the potential impact each threat could have on the FMN infrastructure. This risk-based approach is invaluable in environments where resources and attention must be strategically allocated

to address the most pressing vulnerabilities. By evaluating both the likelihood and impact of a given threat, OWASP allows us to assign risk levels, which are crucial for developing effective mitigation strategies. The ultimate goal is not only to detect vulnerabilities but also to provide actionable insights that cybersecurity teams can use to protect their networks proactively.

The integration of STRIDE and OWASP within our methodology creates a synergy that ensures both a comprehensive identification of potential threats and a prioritized, risk-based approach to addressing them. This dual-layered approach offers several key advantages: it allows for thorough threat modeling while remaining adaptable to the specific needs and configurations of the FMN environment. Furthermore, the methodology is designed with ease of use in mind, ensuring that cybersecurity analysts, regardless of their familiarity with AI technologies, can efficiently implement the methodology. By presenting complex risk assessments in a clear, actionable manner, our methodology empowers organizations to make informed decisions about their AI deployments, safeguarding their data and operational integrity.

Ultimately, this methodology is built to address the unique demands of the FMN context, where the security of AI tools is not just a technical concern but a matter of operational resilience. Our goal is to deliver a practical, user-friendly methodology that not only identifies and mitigates risks but also enhances the overall security posture of the FMN by providing analysts with the tools they need to respond to emerging threats swiftly and effectively. By combining the robustness of STRIDE and the precision of OWASP, we have crafted a risk evaluation methodology that is both comprehensive and adaptable, ensuring that AI-enabled systems can be deployed with confidence in even the most sensitive environments.

4.3 The Full AI Risk Evaluation Methodology

Our AI risk evaluation methodology is designed to systematically assess and mitigate security vulnerabilities associated with AI-enabled tools deployed within FMNs. The methodology is structured around five key steps that ensure a comprehensive evaluation of both the technical and operational risks introduced by these AI systems. Each step is intended to provide detailed insights into the AI tool's security posture, assess potential

vulnerabilities, and guide the development of effective mitigation strategies.

1. **Describe the Role of the AI Tool in the FMN**: The first step in our methodology involves clearly defining the purpose, scope, and functionality of the AI tool within the FMN environment. This includes a thorough understanding of how the AI system integrates into the FMN's existing cybersecurity infrastructure and the specific role it plays in enhancing network defense. For example, is the AI tool used for anomaly detection, threat response automation, or real-time data analysis? By establishing a comprehensive understanding of its operational context, we can better tailor the subsequent risk analysis to the specific tasks and functions of the AI tool. Moreover, this step also requires detailing the communication channels, data flows, and other systems the AI tool interacts with, which is crucial for identifying potential attack surfaces and integration vulnerabilities.

2. **Assess How the AI Tool Could Introduce Vulnerabilities Using STRIDE**: In this step, we systematically evaluate the potential vulnerabilities introduced by the AI tool by categorizing threats using the STRIDE method. Each STRIDE category offers a different lens through which to examine the AI system's susceptibility to attacks. This threat model is especially useful for identifying security gaps in AI tools, which may not have been developed with strong security principles in mind.

    - **Spoofing**: Could the AI system be deceived by fake credentials or manipulated into accepting false identities? This is especially relevant for AI tools that handle authentication or access control, where spoofing attacks could allow unauthorized users to gain access to sensitive FMN systems.
    - **Tampering**: Could an attacker alter the AI system's data inputs, processing models, or outputs? In the FMN context, tampering could lead to inaccurate threat detection or incorrect security decisions, severely undermining the AI tool's efficacy.
    - **Repudiation**: Does the AI system keep sufficient logs and audit trails to ensure accountability? If not, users or attackers could deny having taken specific actions, making it difficult to trace incidents or ensure the integrity of the system's operations. This is a particularly serious concern in a multi-national FMN, where accountability is critical.

- **Information Disclosure**: Could the AI system inadvertently leak sensitive or classified data? In FMNs, AI tools process large volumes of data, some of which could be sensitive. If the system's data-handling protocols are not robust, there's a risk that sensitive information could be exposed, either through model misconfigurations or data breaches.
- **Denial of Service (DoS)**: Could the AI tool be targeted in a Denial of Service attack, rendering it unable to perform its critical functions? AI systems, particularly those that perform real-time analysis or automated response tasks, could be overwhelmed by malicious traffic or excessive inputs, which would significantly degrade their performance or take them offline entirely.
- **Elevation of Privilege**: Could attackers exploit vulnerabilities in the AI system to gain administrative control or higher privileges than intended? This could be catastrophic in the FMN context, where unauthorized access to privileged functions could allow attackers to disable security controls or manipulate the AI system to achieve their own goals.

By systematically mapping the AI tool's functionalities to these STRIDE categories, we gain a holistic view of the system's security posture and potential weak points that need addressing.

3. **Assess the Likelihood of Exploiting Identified Vulnerabilities (OWASP Step 2)**: Once the vulnerabilities have been identified, the next step is to assess the likelihood of each vulnerability being exploited. This involves evaluating the complexity of the attack, the resources and expertise required, and the availability of known exploit methods. Factors such as the attacker's motivation, the system's exposure to external threats, and the sophistication of existing security controls are considered. The likelihood of exploitation is rated on a scale from 1 (very unlikely) to 5 (very likely), allowing for a quantifiable understanding of how realistic each threat scenario is. This assessment is crucial because it helps prioritize which vulnerabilities require immediate attention and which may be less urgent but still need to be addressed in the longer term. A scale of 5 values was chosen because it is adopted for many purposes in different countries (sometimes with a Sixt value representing zero or excellent). To attribute objective metrics to the 1–5 scale for

the likelihood of exploiting a vulnerability the following criteria is proposed in table 4.1:

Scale	Description	Criteria
1	Very Low	• Exploitation requires significant effort or rare conditions. • No public exploits exist. • Only advanced attackers with specific resources could attempt it.
2	Low	• Exploitation is technically challenging but feasible. • Limited availability of public exploits. • Requires a skilled attacker with moderate effort.
3	Moderate	• Exploitation requires moderate effort. • Some public exploits or tools available. • Attackers with common skill levels could succeed under certain conditions.
4	High	• Exploitation is straightforward with minimal effort. • Well-documented or publicly available exploits exist.
5	Very High	• Exploitation is trivial or can be automated. • Commonly used tools or methods can exploit the vulnerability easily.

Table 4.1: Likelihood Scale for Exploiting Vulnerabilities

4. **Evaluate the Impact of Exploiting Vulnerabilities (OWASP Step 3)**: After determining the likelihood of an exploit, we evaluate the potential impact on the FMN if the vulnerability were to be successfully exploited. The impact assessment considers factors such as the scope of the damage, the criticality of the compromised system, and the potential disruption to FMN operations. For instance, a breach in an AI tool used for real-time threat detection could have far-reaching consequences,

including undetected attacks, loss of mission-critical data, or even full system compromise. Impacts are rated on a scale from 1 (minimal impact) to 5 (catastrophic impact), providing a clear sense of how damaging each vulnerability could be. This step is vital for understanding the potential severity of each threat and ensuring that resources are allocated to the most critical risks. It was also evaluated on a scale of 5 values. To attribute objective metrics to the 1–5 scale for the impacts of exploiting a vulnerability the following criteria is proposed in table 4.2:

Scale	**Description**	**Criteria**
1	Very Low	• Negligible impact on confidentiality, integrity, or availability. • Minimal disruption to operations.
2	Low	• Minor impact, such as temporary service degradation. • No long-term harm to organizational goals or data confidentiality.
3	Moderate	• Significant impact on operations or specific systems. • Limited exposure of sensitive data or moderate disruption to services.
4	High	• Severe impact with compromise of sensitive data or critical systems. • Prolonged disruption to services or potential legal/regulatory implications.
5	Very High	• Catastrophic impact, such as full compromise of mission-critical systems. • Major loss of sensitive data, reputational damage, or serious legal/regulatory penalties.

Table 4.2: Impact Scale for Exploiting Vulnerabilities

5. **Prioritize the Risks and Develop Mitigation Strategies (OWASP Step 4)**: The final step involves prioritizing the risks based on the combined likelihood and impact assessments. Vulnerabilities that score high on both scales are deemed high-priority and require immediate mitigation efforts, while lower-priority risks can be addressed over time. Once the risks are prioritized, we develop specific mitigation strategies tailored to the vulnerabilities identified in the earlier steps. These strategies may include enhancing the AI system's security features, implementing additional access controls, improving data encryption, or introducing redundant security measures to ensure system resilience. The aim is not only to patch existing vulnerabilities but also to strengthen the overall security architecture of the FMN in which the AI tool operates.

This structured, five-step approach ensures that all potential vulnerabilities are thoroughly evaluated and addressed. Our AI risk evaluation methodology offers a comprehensive, adaptable, and practical solution for assessing and mitigating the risks posed by AI-enabled tools in FMNs. In these high-stakes environments, where the security of AI systems is paramount, this methodology provides both actionable insights and clear pathways for enhancing overall network resilience.

Moreover, each AI tool has unique functionalities, and their respective roles within the FMN context differ. Therefore, certain STRIDE categories, such as Information Disclosure or Elevation of Privilege, may not always be relevant to every AI tool. For instance, tools that do not handle sensitive data may be less susceptible to Information Disclosure risks. At the same time, those with limited administrative functions may not present significant Elevation of Privilege concerns. By tailoring the methodology to the specific characteristics of each AI tool, we ensure that the evaluation is both precise and relevant to the FMN's operational needs.

4.4 Comparison with Other Risk Management Frameworks

The AI Risk Evaluation Methodology developed in this thesis can be compared to other prominent risk management frameworks discussed in Chapter 3. Each framework provides a structured approach to identifying, assessing, and mitigating risks, but they are

primarily tailored for traditional information systems or web applications. Our methodology adapts and extends these principles to focus on the unique challenges presented by AI-enabled tools in cybersecurity, particularly within the context of FMNs.

The OWASP method provides a risk rating method that emphasizes identifying and prioritizing risks based on likelihood and impact. While our methodology incorporates the OWASP steps for estimating likelihood and impact, the application in AI contexts introduces unique variables. For instance, in AI systems, the likelihood of exploitation is often tied to the complexity and transparency of machine learning algorithms, which OWASP's traditional approach may not adequately capture.

Our methodology extends OWASP by introducing AI-specific considerations, such as data poisoning and algorithmic manipulation, to ensure a more tailored risk evaluation.

The STRIDE method categorizes security threats into distinct categories—Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. However, while STRIDE is commonly applied to general software systems, its focus needs expansion when applied to AI-enabled tools. In the FMN context, AI-driven systems process vast amounts of data and make autonomous decisions, introducing additional risks that STRIDE might not fully address.

For example, AI-specific threats such as adversarial examples (which involve subtly manipulated inputs to fool machine learning models) are unique to AI systems and require additional layers of risk management. Our methodology integrates the STRIDE model but enhances it with AI-centric threat considerations, making it more suitable for evaluating AI tools.

NIST's AI RMF emphasizes a holistic view of managing AI risks, particularly through its GOVERN, MAP, MEASURE, and MANAGE functions. It advocates for a broader perspective, incorporating organizational, ethical, and societal impacts alongside technical risks. While our methodology shares a similar structure in addressing risk identification, measurement, and management, focusing on FMNs requires a more focused, mission-critical application of these principles.

Compared to NIST's broader AI RMF, our methodology is more targeted to the operational and security needs of AI in military networks, where ensuring the confidentiality, integrity, and availability of information is paramount whilst trying to make it as simple as possible. The NIST framework's attention to ethical concerns, such as bias and fairness,

is less relevant in the FMN context, where operational risks, such as system downtime or data breaches, take precedence.

ISO 31000 is a family of standards for risk management, providing guidelines that apply broadly to any type of organization. It emphasizes a structured approach to identifying, assessing, and managing risks while promoting continuous improvement.

However, the AI Risk Evaluation Methodology we have developed is more narrowly focused on AI-specific risks, particularly in high-stakes environments like FMNs. While ISO 31000 advocates for a general risk management process that includes risk identification, analysis, evaluation, and treatment, it does not directly address the unique challenges AI systems introduce, such as adversarial attacks, model biases, or the dynamic nature of machine learning systems.

While our methodology draws on the strengths of these established frameworks, it is designed to better address the specific and evolving risks posed by AI-enabled tools in FMNs. One of the key benefits of our approach is its flexibility, allowing it to consider both AI-specific vulnerabilities and traditional cybersecurity concerns in a straightforward way that SOC teams can easily apply.

4.5 Application of the Methodology to Other Sectors and Organizations

While this methodology has been primarily designed for FMNs in military and governmental contexts, its principles and processes are highly adaptable for other sectors and industries. The methodology provides a versatile foundation for identifying, assessing, and mitigating risks in AI-enabled tools across diverse operational environments.

4.5.1 Adaptability to Critical Infrastructure and Industrial Control Systems (ICS)

Critical infrastructure sectors such as energy, water, transportation, and healthcare increasingly rely on AI-driven systems for real-time monitoring, anomaly detection, and decision-making processes. These sectors face unique challenges, including high operational continuity requirements, sensitivity to service disruptions, and potential human

safety risks. The AI risk evaluation methodology can be applied to assess the vulnerabilities introduced by AI systems in these sectors, particularly in ICS.

- **Spoofing**: In ICS environments, attackers could feed false data into AI-enabled systems that monitor machinery, pipelines, or grid operations, leading to dangerous misinterpretations of normal conditions.
- **Tampering**: Data manipulation in systems that monitor infrastructure components could result in physical damage or service disruption, as AI models may incorrectly classify harmful activity as normal.
- **DoS**: AI tools deployed in critical infrastructure are also vulnerable to denial-of-service attacks, where malicious traffic overloads systems, potentially leading to service outages or safety risks.

By systematically applying this methodology, organizations managing critical infrastructure can proactively identify risks and prioritize mitigation strategies based on the severity and likelihood of potential impacts.

4.5.2 Application in Financial Institutions

Financial institutions are another domain where AI tools are increasingly used, especially in fraud detection, customer behavior analysis, and real-time transaction monitoring. However, these systems are prime targets for cyberattacks that manipulate data or bypass security controls.

- **Spoofing and Tampering**: AI tools that process transactions or flag suspicious behavior may be vulnerable to spoofing attacks where malicious actors feed falsified transaction data, enabling financial fraud to go undetected.
- **Repudiation**: Proper logging and auditing are crucial in financial operations to ensure that any fraudulent activity can be traced back to the source. AI systems that fail to provide these features introduce significant risks in terms of accountability.
- **DoS**: Overloading AI-based fraud detection systems with false positives could cripple their ability to identify real threats, leaving financial institutions vulnerable to unchecked fraudulent activities.

The methodology allows financial institutions to assess their AI systems comprehensively and develop strategies to reinforce operational resilience while ensuring customer trust and regulatory compliance.

4.5.3 Use in Healthcare

AI tools in healthcare are rapidly becoming indispensable for tasks such as medical imaging analysis, patient risk prediction, and clinical decision support systems. However, the sensitivity of personal health information and the critical nature of healthcare operations make the security of these AI systems paramount.

- **Information Disclosure**: AI models processing patient data are particularly vulnerable to privacy breaches. The methodology can be applied to ensure that these systems adhere to stringent data protection regulations such as the *General Data Protection Regulation* (GDPR) in Europe or the *Health Insurance Portability and Accountability Act* (HIPAA) in the United States.

- **Elevation of Privilege**: In healthcare settings, unauthorized access to AI systems controlling clinical decision-making tools could lead to life-threatening consequences if patient data or treatment plans are manipulated.

The methodology can guide healthcare providers in implementing robust security measures that protect sensitive medical data and ensure the integrity of AI-driven healthcare services.

4.5.4 Application to Civilian and Commercial Networks

Civilian and commercial sectors, from e-commerce to logistics, increasingly rely on AI for supply chain management, customer service automation, and marketing analytics. While these sectors may not have the same high-stakes environments as FMNs, they face significant operational, financial, and reputational risks if AI systems are compromised.

- **Tampering**: Manipulating AI models in these sectors can lead to significant business disruptions, such as incorrect inventory forecasting, misaligned marketing efforts, or customer dissatisfaction.

- **DoS**: AI tools automating customer service can be targeted by adversaries, rendering them ineffective and harming customer experience, particularly in high-volume transaction periods such as holiday seasons.

The methodology provides these organizations with a clear and scalable framework to assess vulnerabilities in AI systems and mitigate potential risks, ensuring that AI applications in business processes are secure, reliable, and effective.

Chapter 5

Application of the Methodology

This chapter first applies our five-step AI risk evaluation methodology to OutGene, Cortex XDR, Demisto, and Phantom. We provide a detailed assessment of each tool, including the relevant STRIDE vulnerabilities, separated likelihood and impact assessments categorized by STRIDE, and explanations for excluded STRIDE categories. Afterwards, we will extend this application to the rest of the AI tools referred to in table 3.1 in a more abbreviated way.

By applying our methodology to these tools, we aim to illustrate how the framework systematically assesses vulnerabilities, evaluates the likelihood and impact of potential risks, and prioritizes mitigation strategies in different AI-enabled environments. Combining these case studies will demonstrate the versatility of our evaluation approach and its ability to address a range of AI-related security concerns within FMNs.

5.1 OutGene

Step 1 - Describe the Role of the AI Tool in the FMN

OutGene is an AI-based intrusion detection tool used in FMNs to detect novel or undefined network attacks by clustering host behaviors and focusing on outliers. Its primary role in the FMN context is anomaly detection, which is essential for identifying unknown threats in mission-critical networks. By clustering network behaviors, OutGene identifies patterns that differ from the norm, offering enhanced detection for new or previously unseen attacks that might bypass traditional security measures.

Step 2 - Assess How the AI Tool Could Introduce Vulnerabilities Using STRIDE

The STRIDE methodology allows us to systematically evaluate potential vulnerabilities introduced by OutGene in the FMN environment. Since the tool focuses on network anomaly detection, certain STRIDE categories may apply more significantly than others.

STRIDE Vulnerabilities and Correlation to FMN:

- **Spoofing**: Attackers could manipulate OutGene by feeding it false network behaviors, which in the FMN context could result in OutGene classifying malicious activities as normal, thereby allowing attackers to remain undetected while compromising sensitive mission data.
- **Tampering**: An attacker might tamper with datasets processed by OutGene or interfere with its clustering algorithm. In FMNs, where the tool's ability to detect anomalies is critical for identifying novel attacks, tampered data could result in misclassifying hostile activities as benign.
- **DoS**: Overwhelming OutGene with excessive data could degrade its ability to perform effective anomaly detection. In an FMN context, critical mission systems could be exposed to undetected attacks by overloading the detection tool with unnecessary or fake data traffic.

Step 3 - Assess the Likelihood of Exploiting Identified Vulnerabilities

Following the identification of vulnerabilities using STRIDE, we now assess the likelihood of exploitation in the context of FMN operations.

- **Spoofing**:
  - Likelihood: 3/5 — While spoofing may require specific knowledge of network behaviors, attackers with sufficient insight into FMN traffic could potentially evade detection by feeding manipulated data.
- **Tampering**:
  - Likelihood: 4/5 — Since OutGene relies on data integrity for clustering and detection, tampering is highly likely if attackers gain access to datasets, especially in complex FMN environments where multiple data sources are integrated.

- **DoS**:
  - Likelihood: 4/5 — Attacking OutGene's data processing capabilities with large volumes of fake or erroneous traffic is highly likely, particularly in FMNs where adversaries may aim to overwhelm detection mechanisms.

Step 4 - Evaluate the Impact of Exploiting Vulnerabilities

Next, we evaluate the potential consequences of these vulnerabilities being successfully exploited, focusing on their impact on FMN operations.

- **Spoofing**:
  - Impact: 4/5 — If attackers successfully spoof network behaviors, it could result in malicious activities going undetected, severely compromising FMN operations.
- **Tampering**:
  - Impact: 5/5 — Tampering with data or algorithms could lead to significant misclassification of threats, severely undermining FMN security by allowing adversaries to bypass anomaly detection.
- **DoS**:
  - Impact: 4/5 — A successful DoS attack could disable or degrade OutGene's detection capabilities, leaving FMNs vulnerable to undetected threats during critical mission phases.

Step 5 - Prioritize the Risks and Develop Mitigation Strategies

Based on the combined likelihood and impact assessments, we now prioritize the risks and propose appropriate mitigation strategies for OutGene in the FMN environment.

Mitigation Strategies:

- To counter spoofing, robust authentication mechanisms should be enforced to validate network behaviors and prevent malicious inputs from being treated as usual.
- Data integrity must be preserved by applying digital signatures or encryption on network traffic, ensuring that tampered data can be identified and discarded.

- To prevent DoS attacks, rate-limiting and resource management strategies should be implemented to prevent OutGene from being overwhelmed by excessive traffic.

STRIDE Categories Not Applied:

- **Repudiation**: OutGene does not focus on auditing or action verification, and its role within FMNs does not emphasize tracking individual user actions. Therefore, the risk of repudiation is minimal and is not a priority within this tool's operational scope.
- **Information Disclosure**: OutGene does not handle sensitive or classified data directly, as it primarily works with network flows (metadata). As a result, the risk of information leakage is minimal compared to other systems in FMN that process classified information.
- **Elevation of Privilege**: Since OutGene's role is limited to detection rather than administrative control, the potential for privilege escalation is less relevant.

5.2 Cortex XDR

Step 1 - Describe the Role of the AI Tool in the FMN

Cortex XDR is an extended detection and response (XDR) platform that integrates data from multiple sources to detect and respond to complex, multi-vector threats in real-time. Within the FMN environment, Cortex XDR is crucial for detecting and mitigating advanced persistent threats and other sophisticated attacks that target mission-critical infrastructure. Cortex XDR leverages AI-driven threat intelligence, ML, and behavior analysis to correlate data from various sources and deliver a comprehensive view of potential security risks. This real-time detection capability is essential for FMNs, where rapid responses to emerging threats are critical to safeguarding network integrity.

Step 2 - Assess How the AI Tool Could Introduce Vulnerabilities Using STRIDE

Using the STRIDE methodology, we assess the potential vulnerabilities Cortex XDR could introduce into the FMN, focusing on its multi-source data integration and automated response capabilities.

STRIDE Vulnerabilities and Correlation to FMN:

- **Spoofing**: Attackers could spoof network data or security logs to mislead Cortex XDR into flagging false positives or ignoring legitimate threats. In the context of FMNs, this could result in erroneous responses or failure to detect real attacks, leading to significant disruptions in mission-critical operations.

- **Tampering**: If attackers tamper with the logs, metrics, or configurations that Cortex XDR relies on, they could compromise the tool's ability to detect threats accurately. Tampering could cause Cortex XDR to generate misleading insights, leading to incorrect threat prioritization or inappropriate responses in the FMN environment.

- **DoS**: By overwhelming Cortex XDR with a flood of false positives or irrelevant data, attackers could degrade their ability to detect actual threats, leaving FMN systems exposed to attacks that go unnoticed.

- **Elevation of Privilege**: If attackers gain unauthorized administrative control of Cortex XDR, they could disable its critical detection and response functionalities, leading to a complete compromise of FMN security controls.

Step 3 - Assess the Likelihood of Exploiting Identified Vulnerabilities

- **Spoofing**:
  - Likelihood: 3/5 — Spoofing is moderately likely in FMNs, as it requires a sophisticated understanding of network behavior, but it remains a realistic threat given the complexity of the network environment.

- **Tampering**:
  - Likelihood: 4/5 — The likelihood of tampering is high, especially if attackers gain access to logs or configuration settings. Cortex XDR's reliance on accurate data makes this a high-risk category.

- **DoS**:
  - Likelihood: 4/5 — Overwhelming Cortex XDR with false positives is a likely attack method, especially in environments like FMNs where data flows are complex and high-volume.

- **Elevation of Privilege**:
  - Likelihood: 4/5 — Gaining administrative control over Cortex XDR is a significant risk in FMNs, as it would allow attackers to disable detection mechanisms.

Step 4 - Evaluate the Impact of Exploiting Vulnerabilities

- **Spoofing**:
  - Impact: 4/5 — Successful spoofing could allow attacks to bypass detection entirely, leading to severe disruptions in FMN operations.
- **Tampering**:
  - Impact: 5/5 — Tampering with detection data could have catastrophic effects, leading to undetected attacks that severely compromise FMN security.
- **DoS**:
  - Impact: 4/5 — A DoS attack could incapacitate Cortex XDR's ability to respond to legitimate threats, leaving FMN systems vulnerable during mission-critical events.
- **Elevation of Privilege**:
  - Impact: 5/5 — Unauthorized administrative control could allow attackers to shut down or manipulate detection mechanisms, posing a major security risk to FMN systems.

Step 5 - Prioritize the Risks and Develop Mitigation Strategies

Mitigation Strategies:

- Implement robust access control mechanisms and strict authentication to prevent privilege escalation.
- Apply rate-limiting and anomaly detection to mitigate DoS attacks.
- Ensure data integrity through secure logging and configuration management to prevent tampering.

STRIDE Categories Not Applied:

- **Repudiation**: Cortex XDR implements comprehensive logging, reducing the likelihood of successful repudiation attacks.
- **Information Disclosure**: While Cortex XDR processes sensitive data, its primary function is detection, making the risk of information disclosure relatively low.

5.3 Demisto (Cortex XSOAR)

Step 1 - Describe the Role of the AI Tool in the FMN

Demisto, now integrated in Cortex XSOAR is a SOAR platform that integrates data from various security tools to streamline incident response workflows. Demisto's role is crucial for managing and automating responses to potential security incidents in FMNs. The platform automates repetitive tasks and provides decision-making support for security analysts, enabling quicker and more effective responses to complex security threats. In mission-critical environments like FMNs, where coordination and rapid response are vital, Demisto's ability to automate security processes enhances overall operational efficiency and reduces human error.

Step 2 - Assess How the AI Tool Could Introduce Vulnerabilities Using STRIDE

Demisto's central role in automating security workflows presents certain vulnerabilities, particularly in how it interacts with various integrated security systems. Applying the STRIDE methodology to Demisto helps us understand where risks may emerge and how these risks could be exploited within an FMN environment.

STRIDE Vulnerabilities and Correlation to FMN:

- **Spoofing**: Attackers may attempt to spoof user accounts or sessions, gaining unauthorized access to orchestrated workflows or automated response actions. In an FMN, where rapid and coordinated responses are critical, a successful spoofing attack could lead to unauthorized actions being executed, which may disrupt the entire security apparatus.
- **Tampering**: Given that Demisto integrates data from various systems, an attacker could tamper with the data being processed or the workflows themselves. In an

FMN, where decisions are based on real-time information, tampered data could lead to incorrect responses, delaying critical decisions or even causing false alarms that divert resources away from genuine threats.

- **Repudiation**: Without adequate logging and audit trails, users or attackers could deny actions taken within the platform. This is especially problematic in FMNs, where maintaining accountability and traceability is vital for collaborative operations involving multiple nations or agencies.

- **DoS**: Flooding Demisto with a large number of incidents could overwhelm its automated workflows, impairing its ability to respond to legitimate threats. In an FMN context, where quick incident response is crucial, such a disruption could leave mission-critical systems vulnerable.

Step 3 - Assess the Likelihood of Exploiting Identified Vulnerabilities

Based on the identified vulnerabilities, we assess the likelihood of exploitation for each STRIDE category, focusing on how feasible such an attack would be in the FMN context.

- **Spoofing**:
  - Likelihood: 3/5 — Spoofing requires attackers to gain access to user credentials or sessions, which is moderately difficult but feasible, particularly if authentication mechanisms are weak or misconfigured.

- **Tampering**:
  - Likelihood: 4/5 — Since Demisto integrates with various systems, tampering with its workflows or data is highly likely if attackers can infiltrate any part of the interconnected system. The likelihood is particularly high in FMNs, where multiple systems communicate and data flows are complex.

- **Repudiation**:
  - Likelihood: 3/5 — Without robust logging, there is a moderate likelihood of users or attackers repudiating actions within the platform. In an FMN, this can complicate accountability and investigation efforts.

- **DoS**:
  - Likelihood: 4/5 — Overwhelming the system with numerous fake incidents is a common DoS tactic, and the likelihood is high, especially if Demisto's capacity to handle large volumes of data is exceeded.

Step 4 - Evaluate the Impact of Exploiting Vulnerabilities

We now assess the potential consequences of successfully exploiting the identified vulnerabilities, particularly in terms of their effect on FMN operations.

- **Spoofing**:
  - Impact: 3/5 — Unauthorized access to workflows could lead to incorrect actions being taken within FMNs, which could disrupt coordinated security efforts.
- **Tampering**:
  - Impact: 5/5 — Tampering with security data or workflows could cause widespread disruptions in incident response, leading to severe consequences, such as the failure to mitigate real threats or the misallocation of resources during a crisis.
- **Repudiation**:
  - Impact: 3/5 — The inability to track and verify actions in Demisto could lead to delays in incident response and accountability within FMNs, hampering effective collaboration among allied forces.
- **DoS**:
  - Impact: 4/5 — A DoS attack on Demisto could render the platform unable to respond to genuine incidents, leaving FMNs vulnerable to ongoing or new attacks.

Step 5 - Prioritize the Risks and Develop Mitigation Strategies

Mitigation Strategies:

- Implement strong authentication mechanisms and session management controls to prevent spoofing.
- Ensure data integrity through the use of encryption and secure communications between integrated systems to minimize the risk of tampering.
- Improve logging and auditing to ensure accountability and traceability within the platform.
- Apply rate-limiting and load-balancing techniques to prevent DoS attacks.

STRIDE Categories Not Applied:

- **Information Disclosure**: Demisto focuses on workflow automation and does not directly manage sensitive data, so the risk of information disclosure is minimal in comparison to other tools.
- **Elevation of Privilege**: Since Demisto is more focused on orchestrating security tasks rather than managing privileges, the likelihood of privilege escalation is low.

5.4 Phantom (Splunk SOAR)

Step 1 - Describe the Role of the AI Tool in the FMN

Phantom, now part of Splunk SOAR, is another prominent SOAR platform designed to automate incident response workflows and integrate with a variety of security tools. Within the FMN, Phantom plays a critical role by providing an automated response to threats across different systems. The ability to execute predefined playbooks and coordinate between various security systems allows Phantom to respond to incidents quickly and effectively. This reduces the burden on human operators, allowing them to focus on more strategic decision-making. In the FMN environment, where swift incident response is key to maintaining operational integrity, Phantom's role in automating and streamlining security operations is crucial.

Step 2 - Assess How the AI Tool Could Introduce Vulnerabilities Using STRIDE

Phantom's capacity for integrating and automating responses across multiple systems introduces several vulnerabilities, particularly in how data is processed and how workflows are executed automatically. Using STRIDE, we assess how Phantom could introduce risks to FMNs.

STRIDE Vulnerabilities and Correlation to FMN:

- **Spoofing**: Attackers could spoof inputs or impersonate authorized users to manipulate Phantom's automated workflows. In FMNs, this could lead to unauthorized actions, such as inappropriate responses to incidents or the bypassing of critical security measures.

- **Tampering**: If attackers tamper with Phantom's inputs or workflows, they could disrupt the platform's ability to respond correctly to security threats. In FMNs, such tampering could lead to delays in incident response or misdirected actions that compromise network security.

- **Information Disclosure**: Since Phantom integrates data from multiple systems, improper handling or misconfigurations could result in the unintentional exposure of sensitive FMN data. The risk of information leakage is particularly high if data is shared improperly between integrated systems.

- **DoS**: Phantom could be overwhelmed by excessive automated requests or incident triggers, preventing it from responding to legitimate threats. In an FMN, where real-time responses are crucial, such a disruption could have significant operational consequences.

- **Elevation of Privilege**: Attackers could exploit security flaws in Phantom's access controls to escalate privileges, gaining administrative control over the platform's workflows. In an FMN environment, unauthorized access could lead to system-wide compromises.

Step 3 - Assess the Likelihood of Exploiting Identified Vulnerabilities

We assess how likely it is that attackers could exploit the identified vulnerabilities within Phantom, considering the complexity of each potential attack.

- **Spoofing**:
  - Likelihood: 3/5 — While Phantom's authentication mechanisms are generally strong, spoofing could still be achieved if session management or user authentication is weak.
- **Tampering**:
  - Likelihood: 4/5 — Tampering with inputs or automated workflows is highly likely if attackers can compromise any part of the integrated system.
- **Information Disclosure**:
  - Likelihood: 2/5 — Assuming that Phantom's data is managed securely, the likelihood of information disclosure is relatively low.
- **DoS**:
  - Likelihood: 4/5 — Overloading the system with incident triggers is a common method to execute a DoS attack, making this a highly likely scenario.
- **Elevation of Privilege**:
  - Likelihood: 3/5 — Exploiting access control weaknesses is moderately likely, particularly if permissions and privileges are not adequately managed.

Step 4 - Evaluate the Impact of Exploiting Vulnerabilities

We now evaluate the potential impact of these vulnerabilities, particularly in the mission-critical environment of FMNs.

- **Spoofing**:
  - Impact: 3/5 — Unauthorized actions triggered by spoofed data could lead to incorrect or damaging responses within FMNs, disrupting security operations.
- **Tampering**:
  - Impact: 5/5 — Tampering with workflows could have severe consequences, such as misdirecting resources or delaying responses to actual threats, which could compromise FMN security.

- **Information Disclosure**:
  - Impact: 4/5 — Exposure of sensitive FMN data due to poor configuration or mishandling could have serious consequences, especially if classified information is leaked.
- **DoS**:
  - Impact: 5/5 — A DoS attack could significantly impair Phantom's ability to respond to legitimate threats, leaving FMNs vulnerable during critical incidents.
- **Elevation of Privilege**:
  - Impact: 4/5 — Gaining administrative control over Phantom could allow attackers to manipulate workflows and response mechanisms, potentially undermining the entire FMN security posture.

Step 5 - Prioritize the Risks and Develop Mitigation Strategies

Mitigation Strategies:

- Ensure robust session management and authentication mechanisms to prevent spoofing.
- Implement data integrity measures and secure integration points to prevent tampering.
- Apply rate-limiting and load-balancing strategies to mitigate the risk of DoS attacks.
- Regularly audit permissions and access controls to prevent privilege escalation.

STRIDE Categories Not Applied:

- **Repudiation**: Phantom's logging capabilities ensure that actions are traceable, reducing the likelihood of successful repudiation attacks.

The step-by-step application of our methodology to **OutGene**, **Cortex XDR**, **Demisto**, and **Phantom** demonstrates its effectiveness in identifying, assessing, and mitigating potential risks in AI-enabled cybersecurity tools deployed in FMNs. By leveraging the

STRIDE threat model for comprehensive vulnerability identification and combining it with the **OWASP risk methodology** for structured risk evaluation, we have developed a robust framework that ensures the security and operational resilience of AI systems.

The evaluation of **OutGene** highlighted the strength of our approach in assessing novel AI tools designed to detect undefined network attacks. It systematically uncovered vulnerabilities across multiple threat categories. It provided clear guidance on mitigating high-priority risks, ensuring that AI-driven tools like OutGene can securely deploy in mission-critical environments.

Similarly, applying the methodology to **Cortex XDR** showcased the versatility of the methodology in analyzing more comprehensive AI-enabled platforms. The structured process enabled a clear assessment of the likelihood and impact of potential threats, ensuring the secure deployment of complex, real-time detection tools that integrate diverse security sources.

Extending the evaluation to **Demisto** (Cortex XSOAR) demonstrated the framework's ability to handle SOAR platforms that automate security workflows. The methodology effectively identified risks such as tampering with security workflows and DoS attacks, and it recommended mitigation strategies to ensure the platform's secure integration into FMNs. Demisto's evaluation emphasized the importance of maintaining data integrity and implementing access control mechanisms in environments where automation plays a significant role.

Similarly, the assessment of **Phantom** (Splunk SOAR) further proved the methodology's applicability to security orchestration tools. The methodology identified key risks, including the potential for spoofing or tampering with automated workflows and the risk of information disclosure. By systematically addressing these vulnerabilities, the framework ensured that Phantom could securely automate incident response tasks in the FMN context.

These case studies—OutGene, Cortex XDR, Demisto, and Phantom—demonstrate how our comprehensive AI evaluation framework adapts to different tools and environments within FMNs. It offers a structured, scalable, and practical approach for security analysts, providing the necessary tools to ensure that AI-driven solutions are secure, reliable, and resilient. Ultimately, this methodology is essential for enabling organizations to address evolving cybersecurity threats effectively in high-stakes operational environments.

5.5 Application of Methodology to Remaining AI Tools

In this section, we apply the comprehensive AI evaluation framework to the remaining AI tools listed in Chapter 3: DarkTrace, Vectra AI, Cynet, Cybereason, and Swimlane. These tools 3.1 share various similarities with the four previously studied in-depth (OutGene, Cortex XDR, Demisto, and Phantom). We simplify the methodology application using the same structured process, assessing how these tools could introduce vulnerabilities, estimating the likelihood of their exploitation, evaluating the impact, and prioritizing mitigation strategies.

5.5.1 DarkTrace

Role in FMN: DarkTrace, like OutGene, focuses on real-time anomaly detection using AI. In FMNs, its role is vital for monitoring and identifying threats within segregated networks.

STRIDE Vulnerabilities

- **Spoofing**: DarkTrace could be tricked by falsified network behaviour, leading to the misclassification of legitimate activities as threats.
- **Tampering**: An attacker could manipulate network traffic analyzed by DarkTrace, affecting its anomaly detection.
- **DoS**: Overloading DarkTrace with anomalous traffic could degrade its ability to detect legitimate threats.

Excluded STRIDE Steps: Repudiation and Elevation of Privilege are excluded due to DarkTrace's lack of user or access management functions. Information Disclosure is not a concern since DarkTrace primarily focuses on network traffic analysis rather than handling sensitive data

Likelihood of Exploiting Vulnerabilities

- **Spoofing**: Likelihood: 3 (Moderate).
- **Tampering**: Likelihood: 4 (High).
- **DoS**: Likelihood: 4 (High).

Impact of Exploiting Vulnerabilities

- **Spoofing**: Impact: 4 (High).
- **Tampering**: Impact: 5 (Critical).
- **DoS**: Impact: 4 (High).

Mitigation Strategies

- Implement secure traffic validation protocols and rate-limiting measures to prevent manipulation and overloads.

5.5.2 Vectra AI

Role in FMN: Similar to Cortex XDR, Vectra AI uses behavioral analysis to detect advanced threats in FMNs by monitoring and preventing multi-vector attacks.

STRIDE Vulnerabilities

- **Spoofing**: Attackers may spoof network behavior to evade detection.
- **Tampering**: Like Cortex XDR, tampering with data input or outputs could compromise detection mechanisms.
- **Elevation of Privilege**: Weak access controls could lead to unauthorized escalation of privileges.

Excluded STRIDE Steps: Repudiation and Information Disclosure are excluded as Vectra AI primarily focuses on network behavior, not user data management. DoS is not considered as a critical vulnerability for Vectra AI in the context of FMNs due to the focus on data analysis rather than active traffic handling.

Likelihood of Exploiting Vulnerabilities

- **Spoofing**: Likelihood: 3 (Moderate).
- **Tampering**: Likelihood: 4 (High).
- **Elevation of Privilege**: Likelihood: 3 (Moderate).

Impact of Exploiting Vulnerabilities

- **Spoofing**: Impact: 3 (Moderate).
- **Tampering**: Impact: 5 (Critical).
- **Elevation of Privilege**: Impact: 4 (High).

Mitigation Strategies

- Strengthen access control mechanisms and enforce behavioral baselines.

5.5.3 Cynet

Role in FMN: Cynet autonomously detects and responds to security breaches, similar to Demisto's orchestration capabilities, but with a focus on endpoint protection.

STRIDE Vulnerabilities

- **Spoofing**: Falsified behavior could bypass detection.
- **Tampering**: Tampering with inputs could affect Cynet's response to threats.
- **DoS**: Overloading the system with false alerts could degrade performance.

Excluded STRIDE Steps: Repudiation and Elevation of Privilege are not included due to Cynet's focus on endpoint protection rather than access management. Information Disclosure is not a concern, as Cynet mainly handles network traffic metadata and not sensitive data.

Likelihood of Exploiting Vulnerabilities

- **Spoofing**: Likelihood: 3 (Moderate).
- **Tampering**: Likelihood: 4 (High).
- **DoS**: Likelihood: 4 (High).

Impact of Exploiting Vulnerabilities

- **Spoofing**: Impact: 3 (Moderate).
- **Tampering**: Impact: 5 (Critical).
- **DoS**: Impact: 4 (High).

Mitigation Strategies

- Implement stricter data integrity measures and incident filtering for automation.

5.5.4 Cybereason

Role in FMN: Cybereason uses behavior-based detection to monitor endpoints, similar to Vectra AI and Cortex XDR.

STRIDE Vulnerabilities

- **Repudiation**: Weak logging could prevent accountability for detected incidents.
- **Tampering**: Tampering with endpoint data could lead to incorrect alerts.
- **Elevation of Privilege**: Escalating privileges could compromise system security.

Excluded STRIDE Steps: Spoofing and DoS are less applicable due to Cybereason's endpoint focus rather than network monitoring. Information Disclosure is irrelevant here, as Cybereason primarily focuses on behavioral data, not sensitive information.

Likelihood of Exploiting Vulnerabilities

- **Repudiation**: Likelihood: 3 (Moderate).
- **Tampering**: Likelihood: 4 (High).
- **Elevation of Privilege**: Likelihood: 3 (Moderate).

Impact of Exploiting Vulnerabilities

- **Repudiation**: Impact: 3 (Moderate).
- **Tampering**: Impact: 5 (Critical).
- **Elevation of Privilege**: Impact: 4 (High).

Mitigation Strategies

- Ensure robust logging and privilege management to prevent unauthorized access.

5.5.5 Swimlane

Role in FMN: Swimlane is a SOAR platform that automates incident response, similar to Demisto and Phantom.

STRIDE Vulnerabilities

- **Spoofing**: Spoofing inputs could trigger unauthorized actions.
- **Tampering**: Tampering with workflows or security inputs could disrupt responses.
- **DoS**: Flooding Swimlane with incidents could prevent response to real threats.

Excluded STRIDE Steps: Repudiation and Elevation of Privilege are excluded due to Swimlane's focus on automated incident workflows. Information Disclosure is not a significant concern, as Swimlane handles security operations rather than sensitive data directly.

Likelihood of Exploiting Vulnerabilities

- **Spoofing**: Likelihood: 3 (Moderate).
- **Tampering**: Likelihood: 4 (High).
- **DoS**: Likelihood: 4 (High).

Impact of Exploiting Vulnerabilities

- **Spoofing**: Impact: 3 (Moderate).

- **Tampering**: Impact: 4 (High).

- **DoS**: Impact: 4 (High).

Mitigation Strategies

- Implement rate-limiting and improve input authentication mechanisms.

The AI evaluation methodology, applying the AI risk evaluation methodology, demonstrates that each AI-driven tool faces similar vulnerabilities within FMN contexts. Although each tool has unique features and target functions, the structured approach to identifying spoofing, tampering, DoS, and privilege escalation risks provides a comprehensive analysis for ensuring secure integration into the FMN environment.

The following table summarises the evaluation of each AI-enabled cybersecurity tool based on the methodology developed in this thesis. Each tool has been assessed for vulnerabilities and risks across key parameters such as Spoofing, Tampering, DoS, and Elevation of Privilege. For each parameter, the likelihood and impact are rated on a scale from 1 to 5, and the corresponding risk is calculated by multiplying these values. The risk is further categorized into low, medium, or high to provide a clear understanding of the severity of each vulnerability when deploying AI tools in the FMN environment. This table 5.1 provides a streamlined comparison of the different tools evaluated in the FMN context, facilitating the prioritization of mitigation strategies.

AI Tool	STRIDE	Likelihood (1-5)	Impact (1-5)	Risk (L × I)	Risk Scale
OutGene	Spoofing	3	4	12	Medium
	Tampering	4	5	20	High
	DoS	4	5	20	High
Cortex XDR	Spoofing	3	4	12	Medium
	Tampering	4	5	20	High
	DoS	4	4	16	High
	Elevation of Privilege	4	5	20	High
Demisto	Spoofing	3	3	9	Low
	Tampering	4	5	20	High
	Repudiation	3	3	9	Low
	DoS	4	4	16	High
Phantom	Spoofing	3	3	9	Low
	Tampering	4	5	20	High
	Information Disclosure	2	4	8	Low
	DoS	4	5	20	High
	Elevation of Privilege	3	4	12	Medium
DarkTrace	Spoofing	3	4	12	Medium
	Tampering	4	5	20	High
	DoS	4	4	16	High
Vectra AI	Spoofing	3	3	9	Low
	Tampering	4	5	20	High
	Elevation of Privilege	3	4	12	Medium
Cynet	Spoofing	3	3	9	Low
	Tampering	4	5	20	High
	DoS	4	4	16	High
Cybereason	Repudiation	3	3	9	Low
	Tampering	4	5	20	High
	Elevation of Privilege	3	4	12	Medium
Swimlane	Spoofing	3	3	9	Low
	Tampering	4	4	16	High
	DoS	4	4	16	High

Table 5.1: Likelihood, Impact, Risk, and Risk Scale for Each AI Tool using the AI Risk Evaluation Methodology

In the following graphs, Figure 5.1 illustrates the mean risk value for each AI tool evaluated using our proposed methodology. From this analysis, it becomes evident that **IDS** contribute the highest level of risk to the FMN environment. Conversely, **SOAR** tools are shown to introduce the least amount of risk. This distinction highlights the varying levels of risk associated with different categories of AI-enabled cybersecurity tools, emphasizing the importance of tailored risk management strategies based on the specific functionalities and roles of these tools within FMNs.

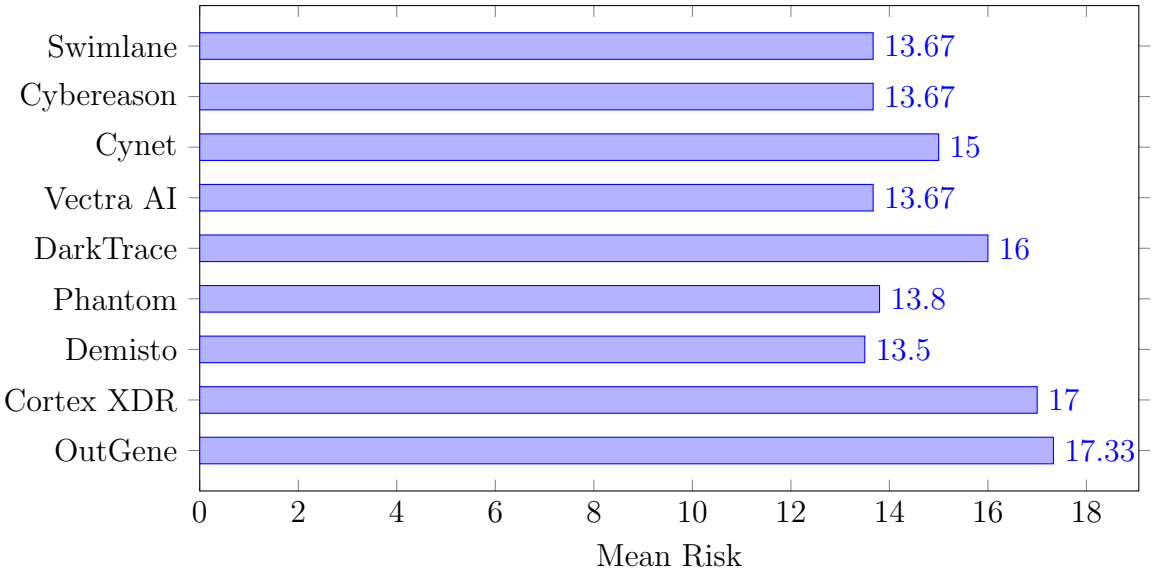

Figure 5.1: Mean Risk Associated with Each AI Tool

Figure 5.2 presents the mean likelihood value for each STRIDE step across all analyzed AI tools. Similarly, Figure 5.3 illustrates the mean impact value associated with each STRIDE step. Finally, Figure 5.4 depicts the mean risk value corresponding to each STRIDE step, providing a comprehensive overview of the assessed risk factors.

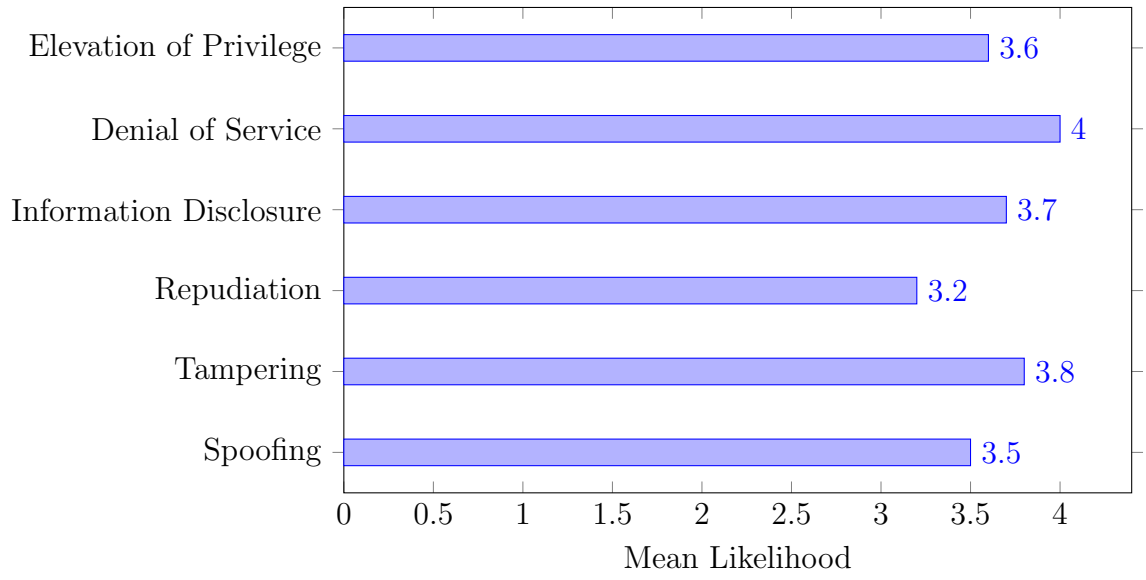

Figure 5.2: Mean Likelihood for Each STRIDE Step

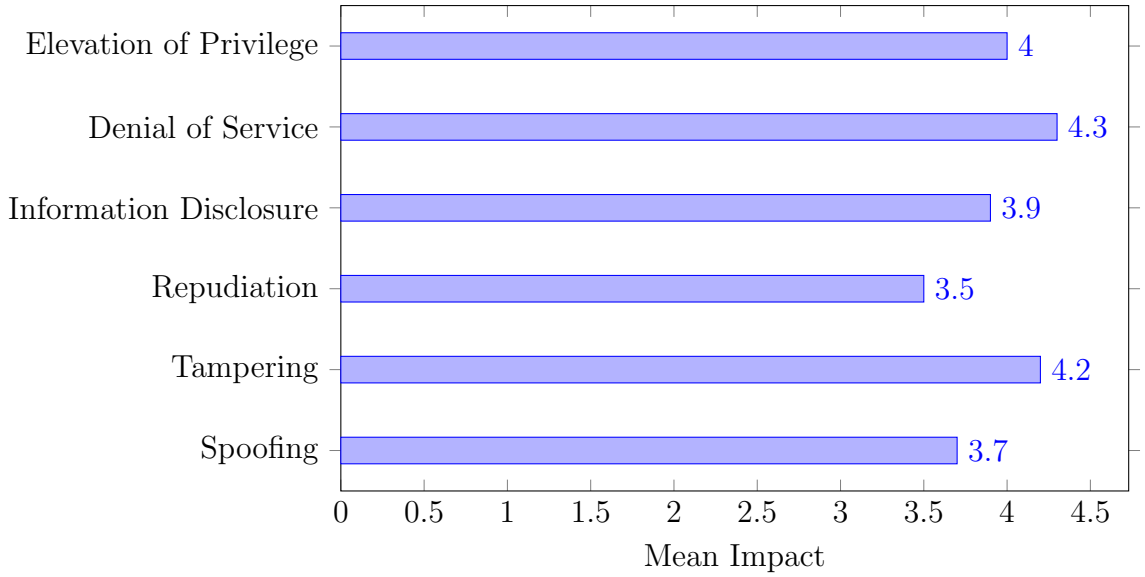

Figure 5.3: Mean Impact for Each STRIDE Step

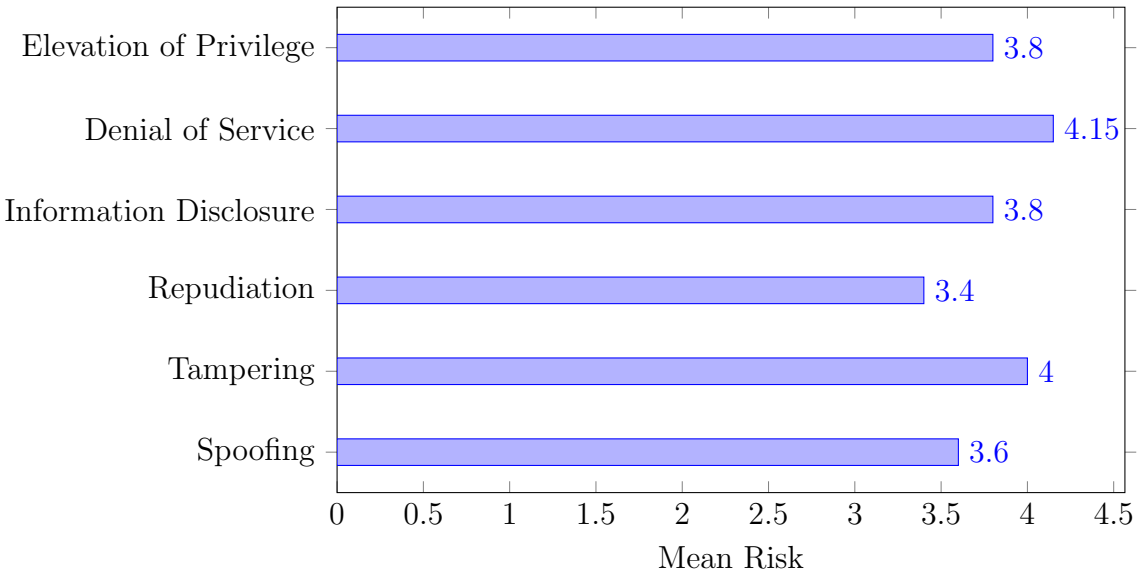

Figure 5.4: Mean Risk for Each STRIDE Step

Through this approach, we can effectively analyze the risks associated with each AI tool, examining the likelihood and impact linked to each STRIDE category. This analysis allows us to identify the STRIDE steps that pose the most significant concern, with **Tampering** and **DoS** emerging as particularly critical risk factors.

Chapter 6

Conclusions

This chapter summarises the key findings and accomplishments of this thesis, highlighting the achievements of the research and outlining potential avenues for future work. The primary goal of this study was to develop a robust and scalable risk evaluation methodology for AI-enabled cybersecurity tools, with a particular focus on FMNs. The methodology was designed to systematically assess the risks associated with using AI in mission-critical environments, leveraging established risk assessment methodologies such as OWASP and STRIDE. Applying the methodology to well-known AI cybersecurity tools demonstrated its practical relevance and effectiveness in addressing real-world security concerns.

The proposed methodology addresses the three questions of the thesis as follows:

- **Q1**: Identifying the critical risks of using AI in cybersecurity applications within Federated Mission Networks (FMNs) was addressed through **Steps 1 and 2** of the methodology. Step 1 involved defining the roles and functionalities of the AI tools within FMNs, while Step 2 systematically evaluated potential vulnerabilities using the STRIDE threat model.

- **Q2**: Evaluating AI tools for their ability to secure FMNs while addressing their vulnerabilities was achieved through **Steps 3 and 4**. Step 3 assessed the likelihood of potential threats by evaluating the conditions and factors that make vulnerabilities exploitable. Step 4 evaluated the impacts of these threats on FMNs, considering their potential consequences on confidentiality, integrity, and availability.

- **Q3**: Ensuring the secure and resilient operation of AI-enabled cybersecurity tools within high-stakes environments was validated through the **implementation of**

the full methodology. This comprehensive application demonstrated the methodology's effectiveness in systematically identifying, evaluating, and mitigating risks in FMNs, enhancing their operational resilience.

This thesis has successfully developed a comprehensive risk evaluation methodology tailored for AI-enabled cybersecurity tools, with a particular focus on FMNs. By incorporating established methods such as OWASP and STRIDE, the methodology effectively identifies, assesses, and prioritizes potential risks associated with the deployment of AI tools in mission-critical environments. Its application to AI-driven cybersecurity solutions like OutGene and Cortex XDR demonstrates its real-world relevance, addressing the increasing demand for robust AI solutions in complex and sensitive network infrastructures.

6.1 Achievements

The development of this framework marks several significant contributions to the field:

The **methodology development** provides a structured and scalable approach for evaluating AI-enabled cybersecurity tools, ensuring adaptability to different contexts and technologies. It incorporates a detailed **risk identification** process, meticulously uncovering AI-specific vulnerabilities and potential threat vectors unique to FMNs. By integrating well-established methodologies with AI-specific considerations, the framework offers a **methodological integration** that bridges traditional risk assessment practices with the nuances of AI technologies. Furthermore, the **practical application** of the framework to prominent AI-enabled cybersecurity tools not only highlights areas requiring improvement but also validates its relevance and utility in addressing real-world cybersecurity challenges.

6.2 Future Work

Building on the foundation established in this research, several avenues for future exploration and development can be pursued:

The framework should evolve in step with emerging AI technologies. This **adaptation for new advancements**—such as deep learning innovations, adversarial AI threats, and AI-based deception techniques—will ensure its continued relevance and effectiveness.

Real-world testing and deployment of the framework in operational environments could further validate its utility and provide invaluable feedback for refinement and iterative improvements.

In addition, exploring the **integration of this framework with other established cybersecurity standards**, such as ISO 27001 or NIST, could broaden its applicability and ensure seamless adoption across different organizational contexts. A critical area for future exploration lies in the **evaluation of adversarial AI risks**, where attackers manipulate AI models to bypass security measures; addressing these vulnerabilities will be crucial for maintaining robust defenses. Finally, the **automation of the evaluation process** through the use of AI or machine learning technologies could significantly streamline the methodology, enabling it to scale effectively for large and complex network environments.

This research has laid a solid foundation for risk evaluation in AI-enabled cybersecurity, providing a pathway for continuous improvement and adaptation in the rapidly evolving landscape of AI and cybersecurity integration.

Bibliography

[1] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, J. Xu, F. Zhang, and K.-K. R. Choo, "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artificial Intelligence Review*, vol. 55, no. 2, p. 1029–1053, Feb 2022.

[2] MITRE Corporation, "Adversarial Tactics, Techniques, and Common Knowledge (ATLAS)," https://atlas.mitre.org/matrices/ATLAS, 2024, accessed: October 2024.

[3] J. Lee and M. Kim, "Machine learning for identification in cybersecurity," *Journal of Machine Learning and Cybersecurity*, vol. 3, no. 1, pp. 45–52, 2020.

[4] P. Nguyen, J. Smith, and A. Gupta, "Deep learning for prevention in cybersecurity," in *Proceedings of the International Conference on Deep Learning and Cybersecurity*, 2019.

[5] A. Ramanathan and S. Wechsler, "Detecting phishing attacks using natural language processing and machine learning," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4561–4569, 2018.

[6] J. Markey, "Using decision tree analysis for intrusion detection: A how-to guide," *SANS Institute*, 2011, Accessed: December 2024. [Online]. Available: https://www.sans.org/white-papers/decision-tree-analysis-intrusion-detection/

[7] W. Eberle and L. Holder, "Insider threat detection using graph-based approaches," in *Proceedings of the IEEE Conference on Intelligence and Security Informatics*. IEEE, 2012, pp. 196–201.

[8] V. Levashenko, F. Abdoldina, V. Gopejenko, K. Yakunin, E. Muhamedijeva, and M. Yelis, "Review of artificial intelligence and machine learning technologies: Clas-

sification, restrictions, opportunities, and challenges," *Mathematics*, vol. 10, no. 15, p. 2552, 2022.

[9] Center for Security and Emerging Technology, M. Musser, and A. Garriott, *Machine Learning and Cybersecurity: Hype and Reality*, Jun 2021, Accessed: December 2024. [Online]. Available: https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/

[10] J.-h. Li, "Cyber security meets artificial intelligence: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 12, p. 1462–1474, Dec 2018.

[11] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, p. 1153–1176, 2016.

[12] N. Oliveira, I. Praça, E. Maia, and O. Sousa, "Intelligent cyber attack detection and classification for network-based intrusion detection systems," *Applied Sciences*, vol. 11, no. 4, p. 1674, Feb 2021.

[13] S. Bhatt, P. K. Manadhata, and L. Zomlot, "The operational role of security information and event management systems," *IEEE Security & Privacy*, vol. 12, no. 5, p. 35–41, Sep 2014.

[14] J. Smith, P. Nguyen, and M. Kim, "Machine learning for security incident and event management systems," in *Proceedings of the International Conference on Machine Learning and Cybersecurity*, 2019.

[15] J. Kinyua and L. Awuah, "AI/ML in security orchestration, automation and response: Future research directions," *Intelligent Automation & Soft Computing*, vol. 28, no. 2, pp. 527–545, 2021.

[16] S. Sarkar, A. Chakraborty, A. Saha, A. Bannerjee, and A. Bose, "Securing air-gapped systems," in *Proceedings of International Ethical Hacking Conference 2019: eHaCON 2019, Kolkata, India*. Springer, 2020, pp. 229–238.

[17] Defence Staff, Communications and Information Technology Directorate and D. Ilie, "Achieving interoperability in a federated environment and in the current security context," *Romanian Military Thinking*, vol. 2022, no. 4, p. 222–235, Dec 2022.

[18] F. T. Johnsen and M. Hauge, "Interoperable, adaptable, information exchange in nato coalition operations," *Journal of Military Studies*, vol. 11, no. 1, pp. 49–62, 2022.

[19] G. Apruzzese, P. Laskov, E. Montes de Oca, W. Mallouli, L. Brdalo Rapa, A. V. Grammatopoulos, and F. Di Franco, "The role of machine learning in cybersecurity," *Digital Threats: Research and Practice*, vol. 4, no. 1, pp. 1–38, 2023.

[20] S. Kaplan and B. J. Garrick, "On the quantitative definition of risk," *Risk Analysis*, vol. 1, no. 1, pp. 11–27, 1981.

[21] OWASP, "Owasp risk rating methodology," https://owasp.org/www-community/OWASP_Risk_Rating_Methodology, Accessed: December 2024.

[22] A. Shostack, *Threat Modeling: Designing for Security.* John Wiley & Sons, 2014.

[23] National Institute of Standards and Technology (NIST), "Risk management framework for information systems and organizations: A system life cycle approach for security and privacy," 2018, [Online; Acessed: December 2024]. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-37r2

[24] ——, *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*, Gaithersburg, MD, Apr 2018, no. NIST CSWP 04162018, [Online; Accessed: December 2024. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

[25] ——, "Security and privacy controls for information systems and organizations," 2020, [Online; Accessed: December 2024]. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-53r5

[26] National Institute of Standards and Technology, "NIST AI 100-1 artificial intelligence risk management framework (AI RMF 1.0)," U.S. Department of Commerce, Gina M. Raimondo, Secretary; National Institute of Standards and Technology, Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology, January 2023.

[27] International Organization for Standardization, "ISO 31000: Risk management – principles and guidelines," ISO, Tech. Rep. ISO 31000:2018, 2018, accessed: December 2024. [Online]. Available: https://www.iso.org/standard/65694.html

[28] "Darktrace," https://www.darktrace.com/, accessed: December 2024.

[29] "Vectra AI," https://www.vectra.ai/, accessed: December 2024.

[30] "CortexXDR," https://www.paloaltonetworks.com/cortex/cortex-xdr-resource-center, accessed: December 2024.

[31] "Cynet," https://www.cynet.com/, accessed: December 2024.

[32] "Cybereason," https://www.cybereason.com/, accessed: December 2024.

[33] L. Dias, H. Reia, R. Neves, and M. Correia, "Outgene: Detecting undefined network attacks with time stretching and genetic zooms," in *Network and System Security*, J. K. Liu and X. Huang, Eds. Cham: Springer International Publishing, 2019, pp. 199–220.

[34] Demisto - Security Orchestration, Automation, and Response Platform. https://apps.paloaltonetworks.com/marketplace/demisto. Accessed: December 2024.

[35] Phantom - Security Automation and Orchestration Platform. https://www.splunk.com/en_us/software/security-orchestration-automation-and-response.html. Splunk. Accessed: December 2024.

[36] Swimlane - Security Orchestration, Automation, and Response (SOAR). https://www.swimlane.com/. Swimlane. Accessed: December 2024.

[37] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge, "The design science research process: A model for producing and presenting information systems research," in *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, pp. 83–106.