Ethical Implementation of AI in Cybersecurity Through Risk-Benefit Analysis



Castro, James

10/28/24

Castro, James

Writing 227z

Rogue Community College

Dear Mrs. Battrick and fellow Writing 227z students:

In the following pages I present my technical report, Ethical Implementation of AI in Cybersecurity Through Risk-Benefit Analysis. Research, writing, and studying this report was a symbol of freedom in my term. Being able to learn about exactly what I wanted was a key factor in my success in this class. Learning about not only cybersecurity but the specifics within has opened opportunity in my career allowing me to broaden my tunnel vision. AI is a massive industry and plays vital roles in cybersecurity. Knowing the purpose and scope allows me to plan for when I solidify my position as a [penetration tester](#_Section_8:_Glossary). I now understand the problems with AI, the ethics, adoption techniques, business practices, solutions, and most importantly, how to work alongside AI.

I am a student majoring in cybersecurity, the field of protection of computer software, systems, and networks from criminal threats focused on unauthorized information theft, hardware sabotage, disruption and misdirection. Since that is my personal mission, this subject is of great importance. I would like to see a world where innocent people and organizations don’t need to fear their personal assets and data be stolen and used for criminal gains. To accomplish this, AI is vital. I now know that without AI, our response time as cybersecurity professionals is slow. We have a fighting chance against security breaches, but with these benefits comes risk. AI also grants specific exploitations for criminals, and accidental privacy violations for professionals. What I learned in this report is how to balance the two sides.

I personally don’t have any experience with AI, nor have I been affected by cybercrime. I have however been targeted by scams, identity theft, and credit card theft, but I was able to see through the attempts and successfully evade disaster. I know Alan Jackson, our expert in cybersecurity affairs at RCC and I have learned much from him. I wrote this report to educate cybersecurity professionals, clients, and stakeholders on the basics of cybersecurity, the benefits and risks of AI, what to consider when implementing AI, and the dangers of cybercrime.

The most important lesson I would like to state is to be aware of the world. For every physical crime, there are about three cybercrimes, according to Statista. Every second of every day, there is someone trying to break into your global bank. Understanding how to protect yourself is invaluable, because you only must protect yourself while the bank must protect millions. Thank you, Mrs. Battrick, for your incredible feedback. I read every word.

Sincerely,

James Castro

Ethical Implementation of AI in Cybersecurity Through Risk-Benefit Analysis

for

Maria Battrick

by

James Castro

December 10, 2024

Table of Contents

[Abstract 5](#_Toc184637793)

[Figure 1: Artificial Intelligence Adoption and Investment Rate Trend 5](#_Toc184637794)

[Section 1: Introduction 6](#_Toc184637795)

[Figure 2: Cybercrime Trends and Projections 7](#_Toc184637796)

[Section 2: Defining the Ethical and Practical Challenges of AI 7](#_Toc184637797)

[AI Overview 7](#_Toc184637798)

[Lack of Moral Guidance and the Violation of Privacy 8](#_Toc184637799)

[Criminal Exploitation 8](#_Toc184637800)

[The Four Keys: Trust, Transparency, Privacy and Accountability 9](#_Toc184637801)

[Section 3: Scope 11](#_Toc184637802)

[Section 4: Methods 11](#_Toc184637803)

[Section 5: Risk-Benefit Analysis of AI in Cybersecurity 12](#_Toc184637804)

[Benefit Analysis 12](#_Toc184637805)

[Enhanced Threat Detection and Response 13](#_Toc184637806)

[Autonomous Vulnerability Detection 13](#_Toc184637807)

[Improved Response Quality 13](#_Toc184637808)

[Malware Detection 14](#_Toc184637809)

[Adaptive Security 15](#_Toc184637810)

[Data Analysis 15](#_Toc184637811)

[False Positive Reduction 16](#_Toc184637812)

[Summary 16](#_Toc184637813)

[Risk Analysis 16](#_Toc184637814)

[Algorithmic Vulnerabilities 17](#_Toc184637815)

[Overreliance on AI 17](#_Toc184637816)

[Concerns with Data Privacy 18](#_Toc184637817)

[Regulation Concerns 18](#_Toc184637818)

[Malicious AI 19](#_Toc184637819)

[Summary 20](#_Toc184637820)

[Section 6: Implementation Recommendations 20](#_Toc184637821)

[Section 7: Conclusion 23](#_Toc184637822)

[Section 8: Glossary 23](#_Toc184637823)

[Section 9: Appendix 24](#_Toc184637824)

[References 25](#_Toc184637825)

**CRTL + Click to Navigate Document Links**

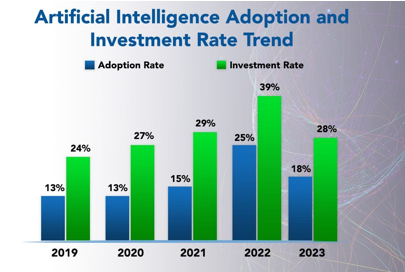
# Abstract

The goal of this report is to explore AI applications and advise readers on implementation techniques through research recommendations. In cybersecurity, AI has reestablished what is possible in terms of response time, efficiency, and general security. In these aspects, AI has the reputation of being indispensable. As AI was being adopted more in the field, new challenges arose that were otherwise unforeseeable. These challenges included new exploitations, unintentional privacy violations, and a lack of moral guidance in AI systems (Cheong, 2024). These critical issues require addressing and will be explored in this document. The TTPA foundations I will be referring to stand for “Trust, Transparency, Privacy and Accountability,” and are the key to successful adoption of AI.

Examining the balance between TTPA allows this study to explore strategies to ethically implement AI in cybersecurity. The methods in this study include an interview with a cybersecurity professor and expert, Alan Jackson, and multiple peer reviewed articles published in journals. Risk-benefit analysis will be performed to establish if AI is indeed as invaluable as it seems; and if so, could the risks be addressed so that AI has less drawbacks? This question and more are answered in this document.

Guidelines for addressing these concerns are found later, ensuring that AI can be implemented ethically, and responsibly. This paper concludes with insights offering guidance to organizations who aspire to utilize AI appropriately, while leveraging its capabilities to secure assets, data, and systems. This report will contribute to the field of cybersecurity with valuable data, clearly stated risk and benefits, guidelines for AI implementation, and acknowledgements for stakeholders, who hold the key to moving the field forward in the ethical direction. Figure 1 demonstrates the effect of stakeholder investments on adoption rates, signaling a dependency on funding.

## Figure 1: Artificial Intelligence Adoption and Investment Rate Trend



Source: Avasant Research: Computer Economics 2023

Section 1: Introduction

In a perfect world, society would be able to exchange their hard-earned finances for goods and services without fear of scam, theft, or sabotage. In reality, any digital exchange, network connection, online account, server and hundreds of other technology driven things or interactions has the potential to be hijacked, stolen, or damaged by cybercrime. Without the proper awareness of cybercrime, people cannot effectively defend themselves against it. Cybercrime costs the world trillions of dollars every year and will only grow as time moves forward (see figure 2, Cybercrime Trends and Projections after introduction), but that doesn’t mean society does nothing to defend itself. Cybersecurity professionals work tirelessly in their careers with software engineers, AI experts and IT personnel to improve their preventative measures and response time to ensure society maintains a safe, digital environment.

There are hundreds of tools at the disposal of cybersecurity professionals. One of the most effective tools in their arsenal is artificial intelligence or AI, as it will be referred to in this document. AI is increasingly utilized in a wide range of situations in cybersecurity. AI is extremely effective in noticing abnormalities and vulnerabilities that can often determine if an attack will, or is occurring, and execute the appropriate alert or countermeasure. AI can detect threats at a greater rate than a person and respond to them even faster. As AI solidified its position as essential, unforeseeable circumstances followed that required investigation and addressing.

The purpose of this document is to address those circumstances and follow with solutions, ideas, and guidance. An example of a problem that arose is the lack of accountability AI holds. AI is designed to be user-friendly and is often seen as a human-like entity. When an issue occurs with the programming and AI fails, or violates client privacy without consent, who to hold responsible can be and has been unclear. This is just an example of one of several concerns this document will discuss.

The heart of this discussion is really to discover how to balance risk and efficiency when implementing AI. Professionals desire the best resources for accomplishing their duties, and when something like AI makes their job easier than ever before, it can be easy to rely far too heavily on it. Addressing the risks and planning for them allows professionals to use this essential tool with complete awareness and take the necessary precautions to negate as many issues as possible. The power of AI can be harnessed but in doing so this field must first master and protect ethics.

## Figure 2: Cybercrime Trends and Projections

A graph of the cost of skyrocket

Description automatically generated

Source: Cybercrime Expected to Skyrocket in Coming Years – Statista

# Section 2: Defining the Ethical and Practical Challenges of AI

## AI Overview

Artificial intelligence has been established as a powerful force in cybersecurity, discovering threats, attacks, and vulnerabilities at unprecedented speeds. Professionals in the field experience this efficiency and commonly develop an overreliance, often allowing AI to perform more complex tasks without the required management that keeps this tool in check. As powerful and useful as AI is, this effectiveness is the same source of ethical and practical challenges that require addressing. With greater implementation of AI, questions of accountability, privacy, overreliance, and exploitation have been increasing in frequency, developing mistrust amongst clients to professionals. As these concerns become more prevalent, a landscape emerges where supervision and management are the key contributors to successful implementation of AI.

## Lack of Moral Guidance and the Violation of Privacy

When AI is given a task, AI does not consider anything other than accomplishing the task. AI is programmed to accomplish that goal without ethical consideration and as a result, AI can violate the privacy of a client by accessing, processing, and memorizing classified resources when only advised to protect them (Taddeo & Floridi, 2018). Storing confidential information in an AI system is an exploitation criminals can take advantage of. This creates logistical and legal challenges where responsibility falls into a grey area. Although these violations are often unintentional, their consequences are significant. The client’s privacy was violated, their data is vulnerable, and there is not a clear responsible party. The AI did as tasked and inspected a resource for vulnerabilities, however as a result of a lack of moral guidance, new vulnerabilities were created and the client lost trust in the professionals (Sargiotis D. 2024). This is commonly referred to as a [security regression](#_Section_8:_Glossary).

## Criminal Exploitation

AI can be exploited in a variety of ways that require addressing. Stated previously, an AI used for security can store confidential information as a privacy violation. Then, it could be stolen on the cybersecurity organizations behalf through improper relegation, storage or shredding of the private data. Another example is by [adversarial attacks](#_Section_8:_Glossary), where cybercriminals manipulate data that’ll be monitored by AI in a way that deceives the system into making incorrect decisions (Malatji & Tolah, 2024). This is done by exploiting vulnerabilities in AI algorithms by altering vital identifying parameters, allowing attackers to bypass security as a misclassified threat. The system believes that the website, file, or even an image is harmless when it is actually malware.

Another common exploitation involves the manipulation of AI systems where an organization considers it autonomous. A system that automatically blocks suspicious IP connections can be exploited through [spoofing. Spoofing](#_Section_8:_Glossary) is done by making legitimate IP addresses appear suspicious, causing the AI system to block them, resulting in critical resources failing to be utilized. Autonomous systems can also be exploited if they are operating completely without human intervention through discovering scenarios AI cannot recognize as an attack pattern. Once discovering this shortcoming, criminals wait for the opportune moment, and then execute a powerful attack that is at this point, unforeseeable. Beyond [adversarial attacks](#_Section_8:_Glossary) and manipulation, other exploitations like [data poisoning](#_Section_8:_Glossary) and [denial-of-service](#_Section_8:_Glossary) exist (Malatji & Tolah, 2024). They are considered more aggressive and an attack of the system itself rather than the resources the system protects.

The greatest tool in a cybercriminal’s arsenal is their version of AI, but instead of defense, it’s designed for attack. This has developed a cyber arms race (Taddeo & Floridi, 2018), where professionals and criminals alike work to develop and counteract advancement in the industry. Exploitations are seemingly limitless, and as a result, cybersecurity professionals cannot afford to waver in developing countermeasures to keep the digital world secure. Human-AI collaboration is the powerful resource the field needs to utilize to ensure a competitive edge in a constantly changing scene.

## The Four Keys: Trust, Transparency, Privacy and Accountability

Effectively using AI in cybersecurity hinges on these four principles: trust, transparency, privacy and accountability. With these elements, an AI system can operate ethically and reliably instilling confidence in clients, operators, and stakeholders. Successful operations depend on all parties feeling confident in the security methods employed. Without a greenlight, the efficiency of protection is hindered resulting in less effective performance and security.

The first element to discuss is trust. Trust between client, professional and stakeholder is the foundation of AI in cybersecurity. AI is relatively new, and some organizations have yet to familiarize themselves with its applications in their digital security. Often, the organization does not inquire about it anyway. Trust lies not between the AI and client, but between the client and the professional managing the AI. If the AI fails, is exploited or results in damage, theft, or sabotage, the professionals and the cybersecurity organizations they belong to lose trust and reputation. This can make AI integration significantly more difficult, especially as it is a new technology and needs to uphold a reputation for reliability and efficiency.

The second element that closely ties into trust is transparency. Transparency is vital in successful implementation, damage mitigation, and customer relations. A cybersecurity professional should always advise clientele on operation planning as it is a symbol of trust and can result in longstanding relations (Cheong, B. C, 2024). The professional should also advise the clientele when a security breach has occurred. If a professional hides this information to protect their company, they can be sued for withholding and will lose all trust with that organization (Cheong, B. C, 2024). Customer relationships are invaluable when exploring options for securing their data and assets. The more trustworthy a cybersecurity organization is, the greater clearance they possess allowing for greater security measures.

The third element is privacy. AI integration requires careful planning and thorough consideration of its role and purpose before usage. It is crucial to ensure that AI does not breach levels of security beyond what is granted by the client (Sargiotis D. 2024). Guidelines for this implementation will be discussed later. Nondisclosures are encouraged when handling private data to instill confidence in the client. If at all possible, access to private data should be avoided entirely if it does not affect security effectiveness. This route ensures that no additional risk is added due to professional negligence (Sargiotis D. 2024). Privacy is the purpose of cybersecurity. To violate it goes against the very nature of the industry.

The fourth and final primary element this document is centered on is accountability. When AI is exploited, violates privacy, or causes other problems, who is held responsible? The AI? The operator? The organization the operator belongs to? This grey area in AI integration remains a significant challenge for the field. Without accountability, responsibility is overlooked and leads to [security regression](#_Section_8:_Glossary) and a lack of correction in critical security errors. According to Cheong (2024), “This lack of transparency and explainability in AI systems poses significant challenges for accountability.” To lay blame on an individual aspect of the vulnerability is to fail to adapt. By understanding all three are accountable, a cybersecurity organization can adapt and evolve while maintaining the trust factor between client and personnel. Legal frameworks often already exist to hold the organization accountable for lack of training and management (Cheong, B. C, 2024). Operators are responsible for misuse or misconfiguration, and while AI is difficult to blame, the developers and systems are not. The AI developers can be contractually liable if their AI system fails to meet performance criteria (Cheong, B. C, 2024). Accountability guides confidence, confidence leads to effective security, and security leads to a safer digital world. To integrate AI is to be accountable.

# Section 3: Scope

This document addresses the challenges of integrating AI into modern cybersecurity operations. Specifically, it examines how the four key principles —trust, transparency, privacy, and accountability—contribute to the ethical implementation. The analysis focuses on a few primary areas in relation to that aspect:

* AI Applications in Cybersecurity
* Risk-Benefit Analysis of AI in Cybersecurity
* Ethics
* Implementation Guidelines

The analysis covers a wide variety of important categories that will define ethical AI integration as a modern standard. The document will cover AI algorithms in the context of cybersecurity including their impact on organizational policies, client relationships and legal frameworks. Narrowing focus to this range of topics provides insights for professionals navigating the complexities of an AI-driven industry.

# Section 4: Methods

To ensure a comprehensive and well-rounded analysis of the challenges both ethical and practical when integrating AI, this document employed two methods of data collection and analysis:

1. **Expert Interview**

An interview was conducted with a seasoned cybersecurity expert with extensive experience in the field. His name is Alan Jackson, and he teaches several networking and security courses at Rogue Community College. He provided valuable insights regarding AI, AI applications, ethics, privacy, work displacement, regulatory framework, and more.

1. **Scientific Articles**

Several scientific articles have been carefully researched and applied to this document. There are many discussions regarding AI in cybersecurity, and several perspectives have been gathered and thoughtfully considered. The research focused on AI vulnerabilities, applications, analytics, ethics, legality, and case studies highlighting successful and unsuccessful AI implementations in cybersecurity.

The research completed for this document provides theoretical and practical foundations that complement the interview with Alan Jackson, ensuring that this document reflects academic rigor and real-world applicability. This dual approach ensures that the guidelines and conclusions are rich in practical experiences and supported by empirical evidence.

# Section 5: Risk-Benefit Analysis of AI in Cybersecurity

This document examines the benefits of AI in cybersecurity, emphasizing both immense potential and the associated risks. While AI enhances detection, prevention, and response, its integration also brings new challenges that require addressing. By identifying these risks, cybersecurity professionals can prepare and mitigate them to ensure AI is used to the full potential, while avoiding new dangerous exploitations (Taddeo, M., McCutcheon, T., & Floridi L. 2019). The ultimate goal of this document is to empower operators and organizations with insight and knowledge to transform unawareness to action, increasing the capacity they possess to secure the digital world.

## Benefit Analysis

Starting with benefit analysis, integrating artificial intelligence into cybersecurity has offered numerous advantages, enhancing the ability to detect, prevent and respond to cyber threats.

Key benefits include:

1. Enhanced Threat Detection and Response
2. Autonomous, Tireless, Vulnerability Hunting
3. Improved Response Quality
4. Malware Detection
5. Adaptive Security
6. Data Analysis
7. [False Positive](#_Section_8:_Glossary) Reduction

### Enhanced Threat Detection and Response

Threat detection and response is a key component in effectively addressing security concerns. It is perhaps the most important because identifying a threat is the first step in prevention, and how quickly a system responds to the threat determines how successful the attack is. AI enables the rapid analysis of complex data, allowing for quick identification of anomalies and potential threats. By using machine learning algorithms, AI can analyze past and current data to detect patterns that indicate imminent cyber-attacks, refining its accuracy with every occurrence. This capability facilities swift response from operators or organizations, ensuring the threat is addressed accordingly based on the severity of alert. For example, AI systems monitoring financial institutions have successfully identified fraudulent activity in real-time by recognizing transaction patterns as unusual.

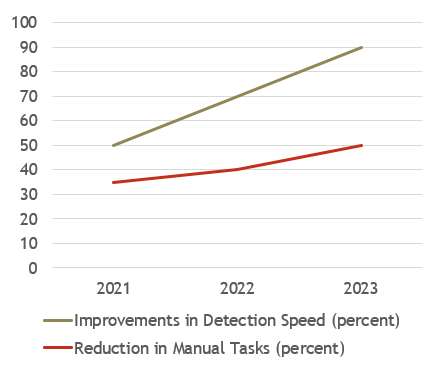
### Autonomous Vulnerability Detection

AI systems are designed to operate tirelessly, outside of human boundaries, performing numerous tasks consistently well, thereby saving cybersecurity professionals time and energy they can allocate to complex tasks. One of its most impactful capabilities is autonomous vulnerability detection. AI-driven systems monitor continuously for potential compromises such as unusual network activity through anomaly detection algorithm utilization, or unauthorized access to secure servers, enabling organizations to address problems before they materialize into cyberattacks. Autonomous vulnerability detection provides personnel with an advantage over criminals, strengthening defenses and allowing for proactive risk mitigation.

### Improved Response Quality

Expanding on automation, AI enhances professional response quality by efficiently completing routine security tasks such as log analysis and alert triage. Alert triage involves categorizing potential threats by severity, ranging from catastrophic to minor. In a traditional system, this task is done by human analysts but can now be automated by AI, where AI uses machine learning models to analyze threat patterns and categorize them based on previous occurrences, granting varying levels of severity. AI streamlines workflows, reducing the burden on professionals allowing them to focus on addressing high-priority threats with greater effectiveness, supported by figure 3 below. In practice, AI-assisted triage has been shown to process thousands of alerts in minutes, ensuring severe threats are identified and resolved, adding to overall crime resilience.

#### Figure 3: Efficiency of AI Trend



*Sources:*

*AI in Cybersecurity statistics for 2024,*

*State of AI and Security Survey Report, CSA*

### Malware Detection

AI models are capable of analyzing software behavior and identifying malicious activity, even from previously unknown sources like new malware variants. This advanced detection technology prevents the spreading of malware before it can dig its roots into a system mitigating potential damage. AI algorithms learn and refine patterns, enabling them to recognize behavior similarities across different malware types. Since malware is typically confined to a limited set of objectives, AI can identify patterns of behavior from known malware variants and apply these insights to detect and quarantine new threats. For example, AI systems have successfully identified [zero-day malware](#_Section_8:_Glossary) by analyzing subtle anomalies in file behavior.

### Adaptive Security

When threats are discovered and addressed, AI systems adapt their security protocols, creating a dynamic defense system that evolves through continuous learning. In addition to dynamic adaptation, AI systems possess global threat monitoring through platforms like [ISACs](#_Section_8:_Glossary) (Information Sharing and Analysis Centers), where organizations share threat data to enhance industry-wide defenses. One limitation of [ISACs](#_Section_8:_Glossary), discussed more later, is their reliance on voluntary contributions; threats are shared only if organizations choose to upload their discovery. In contrast, traditional systems rely on predefined rules and signatures to identify and detect threats, which limit their ability to maintain pace with a rapidly evolving attack scene. By leveraging machine learning algorithms and global intelligence feeds, AI systems hold advantages over traditional systems. This adaptability contributes to the industry’s ability to overtake criminal progression and maintain security for digital assets. For example, an AI system can analyze a new attack pattern shared through [ISACs](#_Section_8:_Glossary) and is then applied to targeted defenses, blocking potential attacks before they can occur.

### Data Analysis

A significant feature of AI is its ability to process and interpret large volumes of data into actionable insights. In cybersecurity, this capability is applied to security data to inform strategic decision-making and strengthen overall security. While the exact amount of time AI saves in analysis varies across studies, the consensus is clear, stating the effect is significant and has shown positive results for the industry. For example, AI systems have been shown to reduce incident response times by analyzing logs and identifying anomalies in minutes, a task that could take human analysts hours or even days to accomplish.

### False Positive Reduction

A [false positive](#_Section_8:_Glossary) is a common occurrence where safe entities are flagged as malware, causing unnecessary interruptions and wasted resources. To address this issue, developments in AI have increased the rate of success when distinguishing between legitimate activity and potential threats. For instance, a study presented at the [4th Workshop on Artificial Intelligence-Enabled Cybersecurity Analytics](https://dl.acm.org/doi/10.1145/3637528.3671494) demonstrated that a machine learning framework could suppress 54% of [false positives](#_Section_8:_Glossary) while maintaining a 95.1% detection rate. In practice, cybersecurity professionals are able to prioritize real threats consistently thanks to this development, saving time and resources for maintaining secure digital environments.

### Summary

The integration of AI into cybersecurity frameworks has transformed the industry, significantly increasing an organization’s ability to defend against evolving cyberthreats. Through advancements in threat detection, adaptive security, and data analysis, AI has been proven to save organizations time and valuable resources for complex challenges and threats that require human intervention. AI continues to build resilient foundations in cybersecurity infrastructure, making it increasingly difficult for cybercriminals to achieve their objectives. As Alan Jackson noted in an interview, “AI is in its infancy,” suggesting it has great potential for future advancements.

## Risk Analysis

The integration of artificial intelligence into cybersecurity brings unparalleled benefits, but it also introduces risks that require careful analysis for proper mitigation (Taddeo, M., McCutcheon, T., & Floridi L. 2019).

Key risks include:

1. [Algorithmic Vulnerabilities](#_Section_8:_Glossary)
2. Overreliance on AI
3. Concerns with Data Privacy
4. Regulation Concerns
5. [Malicious AI](#_Section_8:_Glossary)

### Algorithmic Vulnerabilities

AI systems are only as secure as the algorithms that drive them, and if vulnerabilities exist in the machine learning models or algorithms, attackers can exploit them to bypass security measures. The first example is [adversarial attacks](#_Section_8:_Glossary), where attackers manipulate input data to mislead AI systems (Malatji, M., & Tolah, A, 2024). This causes incorrect threat identification resulting in more [false positives](#_Section_8:_Glossary) (Taddeo & Floridi, 2016) and less correct detections. The second example is poisoning attacks. These attacks introduce corrupt or misleading data that AI uses to adapt through machine learning, so the AI holds biases or flaws in future predictions. The result of this vulnerability is a compromised system that is easier to penetrate. Along with easier penetration, there is less trust in AI solutions from cybersecurity professionals, clients, and stakeholders alike. To address these vulnerabilities, operators should regularly update and test AI models. [Penetration testing](#_Section_8:_Glossary) involves simulating attacks from a cybersecurity professional’s perspective. This hands-on method is an effective way to discover critical vulnerabilities and patch them in future updates. To mitigate poison attacks, AI developers should adopt robust adversarial training techniques to prepare models for manipulations through understanding and ignoring suspicious datasets. Adversarial training involves controlled machine learning environments where AI is exposed to manipulative datasets and rewarded for correctly identifying and discarding them.

### Overreliance on AI

AI systems are not miracles and require human oversight to ensure operations happen smoothly and without error. Because AI systems can be so efficient, professionals have become over reliant, often allowing AI to perform mundane tasks indefinitely without supervision or updates. When an organization relies exclusively on AI, nuanced or contextual threats that require human intuition can be missed. Alerts flagged by AI as “low risk” are often ignored by these organizations as they trust AI abilities too fiercely. When a “low risk” threat is falsely classified, it is a critical vulnerability that can potentially become a security breach. Relying entirely on autonomous systems results in missed opportunities for effective threat remediation (Taddeo et al., 2023). To fix this, professionals must maintain a balance between AI automation and human supervision.

### Concerns with Data Privacy

AI systems often require vast amounts of data for training and operations. This raises concerns about the handling and storage of sensitive information. AI systems have inadvertently stored or shared private client data on occasion, resulting in a critical violation of their privacy and a violation of contractual agreement. Such incidents also violate data protection regulations such as [GDPR](#_Section_8:_Glossary) (General Data Protection Regulation). This results in a loss of client trust and reputational damage. It also results in legal and financial penalties for non-compliance if the mistake is not reprimanded. Finally, new vulnerabilities are created when AI improperly manages data, providing additional opportunities to cybercriminals. To mitigate these risks, [federated learning](#_Section_8:_Glossary) offers a promising solution. This machine learning approach allows multiple organizations to collaboratively train a single AI model without sharing raw data, prioritizing privacy and minimizing risks. [Federated learning](#_Section_8:_Glossary) ensures the AI system operates independently while avoiding storage of sensitive information. Additionally, operators and AI developers can implement compliance protocols to prevent the AI from inadvertently processing private data, ensuring legal and ethical compliances. Together, these strategies reduce the risk of privacy violations and strengthen trust in AI.

### Regulation Concerns

Regulations for AI in cybersecurity are not as clearly defined as they should be, particularly regarding accountability. This creates concerns for the clientele, especially when they are eligible for legal compensation. For instance, when an AI system fails to detect a threat, resulting in stolen or damaged assets, determining liability remains unclear. As discussed in the "Four Keys" section earlier in this document, assigning responsibility among AI developers, operators, and organizations is a pressing issue that requires federal-level attention.

Privacy is another regulated area in cybersecurity, more so than accountability, yet AI systems have occasionally violated these standards. Cybersecurity organizations must ensure their AI systems comply with regulations such as [GDPR](#_Section_8:_Glossary) (General Data Protection Regulation) and the EU artificial intelligence act (Floridi et al., 2023) or face consequences. Maintaining transparency, addressing privacy concerns and ensuring legal standards are met with is vital for instilling confidence in clients (Taddeo, M., & Glorioso, L, 2016).

Overreliance on AI systems also warrants stricter regulation. Operators should be legally required to conduct periodic audits of their systems to ensure that AI makes reliable and unbiased decisions. These audits would provide oversight and help organizations maintain trust with their clients while adhering to ethical and legal standards. “Algorithmic auditing and impact assessments are vital for enhancing accountability in AI systems.” Kroll et al. (2017). Together, clear accountability frameworks, robust privacy safeguards, and regular audits can address critical regulatory gaps and ensure the responsible use of AI in cybersecurity.

### Malicious AI

AI systems are constantly evolving, improving threat detection and response times. While these advancements benefit the industry, they also pose significant risks when accessed and manipulated by cybercriminals (Taddeo, M., Mccutcheon, T., & Floridi L. 2019). A cyber arms race is actively unfolding between cybersecurity professionals and cybercriminals, both striving for a more powerful AI system than the other (Taddeo & Floridi, 2018).

As powerful AI is for defense, it is equally potent for attack. Both types of AI systems operate similarly, tirelessly searching for vulnerabilities. The defining difference lies in their end goals: one aims to secure systems and protect assets, while the other seeks to exploit weakness for gain. For instance, cybercriminals leverage AI to automate [phishing](#_Section_8:_Glossary), where a criminal deceives members of an organization to send them critical information (Malatji, M., & Tolah, A, 2024). They possess the ability to automatically generate email addresses similar to a legitimate organization’s, generate a hyper-realistic email requesting sensitive data, and then even send it to a recipient. Another example is automated malware distribution, where AI systems adapt to evade cybersecurity measures based on pattern recognition. Regardless of the method, [malicious AI](#_Section_8:_Glossary) systems illustrate the urgency to maintain robust defense to stay ahead of the arms race.

### Summary

The risks of AI are synonymous with inefficiency, as it is required by organizations to address risks if they are to maximize AI potential. Thorough analysis of risks enables professionals to address them, mitigating potential legal problems, disruption, exploitation, and vulnerabilities. Regulations provide a critical framework for maintaining industry standards and client confidence. Adhering to regulatory requirements ensures professionals uphold their responsibility to address threats, avoid privacy violations, all while not creating new vulnerabilities criminals may exploit (Taddeo, M., & Floridi L. 2017). AI technology will continue to evolve, and as it does, professionals must stay vigilant addressing these challenges so that they may maintain trust and security in the digital age.

# Section 6: Implementation Recommendations

Cybersecurity is a field where AI has become indispensable, and as a result the necessity for organizations to address risks and concerns has grown. If each risk is effectively addressed, AI systems are capable of significantly improving the industry security standards. Only by thoughtful consideration can AI be effectively utilized, so in this section this document will advise individuals on the most likely to succeed methods to implement AI while effectively addressing vulnerabilities, ethical concerns, and exploitations.

1. Establish Accountability Framework
   1. Purpose: Instill confidence in professionals, clients, and stakeholders.
   2. Guidelines:
      1. Define accountability for AI developers, operators, and cybersecurity organizations.
      2. Employ [SLAs](#_Section_8:_Glossary) (Service Level Agreements) to specify responsibilities and expectations for AI systems.
      3. Regularly update, test and document AI systems to ensure compliance and transparency standards are met.
2. Guarantee Data Privacy
   1. Purpose: Protect sensitive data, assets, and information by complying with [GDPR](#_Section_8:_Glossary) regulations and industry standards.
   2. Guidelines:
      1. Implement techniques that preserve privacy such as [federated learning](#_Section_8:_Glossary) to train AI without compromising security.
      2. Develop strict access controls and encryption protocols to prevent unauthorized access to private data.
      3. Regularly audit AI systems for inadvertent data recording and compliance with data protection laws local and international.
3. Conduct [Bias Audits](#_Section_8:_Glossary)
   1. Purpose: Minimize training biases in AI systems to ensure AI decisions are made fairly and appropriately, especially without attacker manipulation.
   2. Guidelines:
      1. Regularly test AI models for biases, using diverse datasets to increase inclusivity.
      2. Establish a dedicated team for reviewing and correcting bias-related concerns.
4. Balance Automation and Human Oversight
   1. Purpose: Address overreliance concerns in organizations.
   2. Guidelines:
      1. Implement AI to automate routine tasks such as log analysis and alert triage to allow humans to focus on complex tasks, but regularly inspect AI decisions to ensure algorithmic flaws are not present.
      2. Train cybersecurity professionals to interpret and validate AI output, such as threats flagged as low risk during alert triage, in case the AI system is flawed.
      3. Regulate automation and enforce human oversight to ensure threats that require human intuition are addressed.
5. Mitigate [Algorithmic Vulnerabilities](#_Section_8:_Glossary)
   1. Purpose: Strengthen AI systems against attacks like adversarial and poisoning attacks.
   2. Guidelines:
      1. Train AI by exposing it to manipulative data sets, rewarding for successful identification.
      2. Regularly perform [penetration testing](#_Section_8:_Glossary) on AI models to identify and patch vulnerabilities.
      3. Update algorithms frequently for adaptation purposes.
6. Establish Transparency Framework
   1. Purpose: Build trust amongst clients, stakeholders, and professionals by recording and interpreting AI decisions into easy-to-understand information.
   2. Guidelines:
      1. Adopt AI technology that provides clear insight for autonomous decisions.
      2. Document and share decision-making processes and system limitations with clients.
      3. Create visual tools to illustrate AI performance.
7. Monitor AI Systems
   1. Purpose: Maintain system security and accuracy for future operations.
   2. Guidelines:
      1. Establish a time interval for review cycles to assess AI performance.
      2. Search for changes in data patterns that may affect accuracy, retraining models as required.
      3. Collaborate with platforms like [ISACs](#_Section_8:_Glossary) to integrate the latest threats into AI systems.
8. Promote Ethical Standards
   1. Purpose: Align AI systems with organizational values and societal expectations.
   2. Guidelines:
      1. Implement and enforce organizational policies to address accountability, fairness, and privacy.
      2. Involve ethicists and legal experts when developing or deploying AI systems.
      3. Clearly define responsibilities and commitments to clients and stakeholders and uphold them.

# Section 7: Conclusion

The integration of AI into cybersecurity has undeniably benefited the industry, changing how professionals detect threats, analyze data and respond to crises. As useful as the iterations are, there comes the risk that requires appropriate addressing. Ethical concerns surrounding categories like accountability, bias and overreliance highlight the importance of careful implementation, addressing and mitigating risks to maximize AI potential.

By addressing AI challenges, organizations can be confident and aware during use. AI is a powerful tool used to secure the digital world but requires transparency and trust to be considered to maintain client blessings. Regulatory frameworks require greater presence for clarity purposes, preserving client faith in their cybersecurity professionals. Clear accountability, privacy centered practice, regular audits and ethical standards are essential to ensure AI systems operate responsibly and effectively.

As AI continues to evolve, its presence in the industry will only become more formidable. Cybersecurity professionals, clients, and stakeholders collectively are what shapes this formidable force. It is the responsibility of all three to ensure it is shaped ethically and securely for future operations. Through vigilance, innovation and adherence to ethical principles, the industry can harness AI’s potential while maintaining the integrity of the digital environment.

This document hopefully served its ultimate purpose to convey the benefits and risks of AI while supplementing basic implementation guidelines. Understanding is the key to evolution, and this document advocates for that awareness. Only by understanding AI can it be truly utilized. It is recommended for individuals interested in cybersecurity, or specifically AI in cybersecurity, to attend events such as the [Workshop on Artificial Intelligence-Enabled Cybersecurity Analytics](https://dl.acm.org/doi/10.1145/3637528.3671494) to learn more. Readers may also consider viewing the sources in this document to develop a thorough understanding of the topic. I recommend professionals in the field develop their understanding of AI so that they may utilize it with confidence. Cybersecurity is the primary defense against digital theft and sabotage, so cybersecurity professionals who may read this document, continue to develop insight to develop advantages against criminals doing the same.

# Section 8: Glossary

* **Adversarial Attack:** A method of attack where data is manipulated to deceive AI systems, resulting in incorrect classifications of threats and malware.
* **Algorithmic Vulnerabilities:** A weakness in the algorithms that drive AI that can be used by attackers to compromise a system.
* **Bias Audit:** The evaluation of AI systems to identify unintended preferences or bias in AI systems often as a result from data poising.
* **Data Poisoning:** An attack where training data is compromised and manipulated into causing AI to develop flawed or biased learning outcomes.
* **Denial-of-Service Attack:** An attack that disrupts a user or organization’s ability to access/provide resources, usually by overwhelming their system with excessive data or requests.
* **Federated Learning:** An AI training method where multiple entities collaborate together without sharing raw data.
* **False Positive:** An incorrect threat classification on an innocent entity
* **GDPR (General Data Protection Regulation):** A regulatory framework designed by the EU to protect personal data and privacy for individuals in Europe.
* **ISAC (Information Sharing and Analysis Center):** An organization where members share threat intelligence in order to collaboratively strengthen defenses against cybercrime.
* **Malicious AI:** Artificial intelligence designed to perform harmful tasks, such as spreading malware or phishing.
* **Penetration Testing:** The act of cybersecurity professionals attempting to exploit previously unknown vulnerabilities and correct them.
* **Phishing:** A scam or attack where criminals deceive individuals into sharing sensitive data, ranging from company secrets to credit card information.
* **Security Regression:** When a system or professional implements new security measures that inadvertently create more vulnerabilities than are addressed.
* **SLA (Service Level Agreement):** A contract between client and professional that defines responsibilities and expectations.
* **Spoofing:** An attack where criminals disguise their identity by falsifying data such as email addresses or IP addresses to appear legitimate. Spoofing can also occur when a criminal disguises legitimate resources as a threat to block systems from receiving necessary data.
* **Zero-Day Malware:** Newly developed malware deployed by criminals to attack a system resulting in unprepared defenders without countermeasures.

# Section 9: Appendices

## Appendix A Full Alan Jackson Interview

*Q: How would you define "ethical implementation" of AI and what criteria should it meet?*

A: This is a very difficult and broad topic to discuss. AI has been around in the cyber security world for a long time. For example, I would suggest that heuristic scans in antivirus could qualify as AI in cyber security. The biggest ethical concerns about AI in cyber security deal with decision making. Computers don’t use ethical decision-making frameworks unless they are specifically told to do so. However, the nature of AI necessitates that AI make decisions in the abstract. If the AI is learning to make better decisions, it is learning on its own. And the algorithm programmers themselves may not know how a computer system came to a conclusion. Therefore, I would suggest that there should always be a human who looks at decisions being made by an AI system. This human must determine if the decisions should be followed.

This would necessitate the human to clearly define and codify the ethical framework that he/she will use to make a determination. In this scenario, the AI would make recommendations, and the person would ratify them. The ultimate liability will still be on the individual, rather than the AI that made the recommendation in the first place.

Q: *What does AI bring to the table in terms of risk detection, response time, and efficiency?*

A: This could be very beneficial for AI professionals. One of the biggest challenges facing businesses is simply detecting malicious behavior. Setting up tools to parse through hundreds of thousands of logs on every conceivable system has always been very difficult and time consuming. Furthermore, these systems require constant tuning and tweaking to keep from getting flooded with [false positives](#_Section_8:_Glossary). If an AI system can be implemented with the knowledge of what to look for, it could drastically improve risk detection, response time, and efficiency. A next-gen IDS/IPS system or SEIM would HAVE to be substantially better than old school systems.

Q: *Follow up question: Do you see it evolving or dissolving as time moves forward?*

A: This will only improve over time. Currently SEIM systems cost so much money that only large companies can afford them. Furthermore, the expertise to run them is hard to come by. I hope that SEIM, IDS/IPS, and others become so commoditized that small business can subscribe to a cloud service for cheap and have their security covered.

Q: *What are the greatest risks or ethical concerns when using AI, particularly regarding privacy and data protection?*

A: Again… big question. You must ask yourself who owns the AI? Does the company have a particular bias? Where does the data go? And if companies are sending security information into a system to be analyzed, how does that system secure it? Hypothetically, you could have an AI cyber security system that is designed to look at Windows logs to look for malicious behavior. However, those logs also contain various private bits of information that a company may not want disclosed. Once it is run through the AI system, does it belong to the company anymore? You are training a system when you feed it information so all parties benefit from the training. I am not sure how we answer that question. And it is complicated by the fact that foreign countries have much stricter laws on data privacy than the U.S. does. How do we work with them?

Q: *How could AI potentially cause job displacement in the field?*

A: Well if AI systems were designed to detect malicious behavior, and there was no human checking or validating the results, then this could displace InfoSec workers. However, a business would then run the risk of making unethical decisions based on a computer’s recommendation.

Q: *Do you have any suggestions how to mitigate job loss while retaining efficiency?*

A: There are so many cyber security jobs that are currently unfilled in the IT world, that I am not worried about job loss in the near future. In fact, I think that this technology could backfill a large need. With that said, the cyber security professional’s job may have to morph into more of a liaison role. Employees may now need to take reports from AI, digest them, and explain them to non-technical people who don’t understand the implications of the AI report. I don’t see the need for InfoSec going away. It may just be different.

Q: *Are you familiar with any real-world examples of ethical implementation?*

A: Not of AI in the cyber security world. I have seen ethical implementations of AI to address other business needs. I think that AI in cyber security is still in its infancy.

Q: *Is our current regulatory framework sufficient, or are additional regulations necessary to ensure appropriate AI usage?*

A: Our regulatory framework is never sufficient to address issues like this. I have worked in many highly regulated industries where cyber security is mandated. However, businesses who value cyber security implement it. Businesses who don’t value it look for ways to avoid it. If the business who avoids it is required to have it, they will only do enough to avoid penalty. However, you can’t force someone to feel the need. In many cases, the only thing that changes a company’s mindset is getting breached. Fortunately, or unfortunately, that is happening all the time. Some day we will wake up and put the systems in place to keep us more secure. Once again… AI systems that become so commoditized that anybody can afford them, will bring about a huge improvement to the cyber security world.

## Appendix B Workforce Impact Insight:

Workforce impact was discussed in previous versions in greater detail, but after the interview with Alan Jackson, he provided transforming information that encouraged scrapping of the section.

*“There are so many cyber security jobs that are currently unfilled in the IT world, that I am not worried about job loss in the near future. In fact, I think that this technology could fulfill a large need. With that said, the cyber security professional’s job may have to morph into more of a liaison role. Employees may now need to take reports from AI, digest them, and explain them to non-technical people who don’t understand the implications of the AI report. I don’t see the need for InfoSec going away. It may just be different.”*

To interpret, he basically says the industry need for professionals is so significant, that the effect of AI on job retention is minimal if not none, but instead suggests AI will alter the responsibilities of professionals.

# References

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. AI and Ethics, 4(1), 27–44. <https://doi.org/10.1007/s43681-024-00427-4>

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6, Article 1421273. <https://doi.org/10.3389/fhumd.2024.1421273>

Sargiotis, D. (2024). Ethical AI in information technology: Navigating bias, privacy, transparency, and accountability. Social Science Research Network. <https://doi.org/10.2139/ssrn.123674069>

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. Nature, 556(7701), 296–298. <https://doi.org/10.1038/d41586-018-04602-6>

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. Nature Machine Intelligence, 1(12), 557–560. <https://doi.org/10.1038/s42256-019-0109-1>

Taddeo, M., & Glorioso, L. (Eds.). (2016). Ethics and policies for cyber operations: A NATO cooperative cyber defence centre of excellence initiative. Springer. <https://doi.org/10.1007/978-3-319-47852-4>

Taddeo, M., & Floridi, L. (2017). The responsibilities of online service providers. Minds and Machines, 27(2), 265–268. <https://doi.org/10.1007/s11023-017-9436-3>

Floridi, L., Holweg, M., Taddeo, M., Silva, J., Mökander, J., & Wen, Y. (2023). CapAI - A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. Oxford Internet Institute and Saïd Business School, University of Oxford. <https://doi.org/10.2139/ssrn.1234567>

Cloud Security Alliance. (2024). *The state of AI and security survey report*. Retrieved December 10, 2024, from <https://cloudsecurityalliance.org/artifacts/the-state-of-ai-and-security-survey-report>

Statista. (n.d.). *Expected cost of cybercrime until 2027*. Retrieved December 10, 2024, from <https://www.statista.com/chart/28878/expected-cost-of-cybercrime-until-2027/>