Securing Artificial Intelligence Systems: Adapting Traditional Cybersecurity to a New Domain

Michael Lively-Scholz
4/21/26

Executive Summary

As artificial intelligence (AI) systems become increasingly integrated into critical sectors such as healthcare, finance, and national security, concerns surrounding their security have grown significantly. Traditional cybersecurity measures, while effective for conventional systems, are not fully equipped to address the unique vulnerabilities associated with AI technologies. As AI continues to develop into its own domain we begin to enter uncharted territory. With this new domain comes new vulnerabilities, such as data poisoning, adversarial attacks, and model exploitation. These each pose serious risks to data integrity, privacy, and system reliability. Beyond traditional cybersecurity techniques, we need to adapt our security mindset to account for and protect these vulnerabilities associated with artificial intelligence; as the artificial intelligence domain emerges and quickly evolves, we must also quickly evolve by establishing a new field: AI Security.

This paper argues that existing cybersecurity approaches are insufficient for protecting AI systems and highlights the need for specialized security strategies. It also establishes the potent idea that frameworks such as the National Institute of Standards and Technology (NIST) AI Risk Management Framework 1.0 (AI RMF 1.0) represent important progress, they contain notable gaps that limit their effectiveness in addressing technical AI-specific threats. This paper outlines key risks, analyzes current challenges, and proposes strategies for improving AI security moving forward.

---

1. Introduction

Artificial intelligence is rapidly transforming industries by enabling automation, enhancing data-driven decision-making, and implementing advanced predictive capabilities directly into workflows. From autonomous vehicles to medical diagnostics, AI systems are

increasingly relied upon in high-stakes environments. However, as the adoption of AI expands, so do concerns about the security of these systems.

Unlike traditional software, AI systems accept complex prompts and output information that is influenced by large datasets and complex models, making them vulnerable to unique attack vectors, not commonly seen across other systems. These vulnerabilities are not fully addressed by commonplace cybersecurity practices, which were not designed with AI-specific risks in mind. To continue an environment of robust security, we must quickly and efficiently implement new practices, while ensuring we do not do so in haste. As a community, we must work diligently in collaboration to design new standards that will act as a baseline for decades to come.

This paper examines the emerging security challenges associated with AI systems, including key threat vectors and limitations of current approaches. It also evaluates the effectiveness of existing frameworks and proposes strategies to improve AI security.

Thesis:

Traditional cybersecurity approaches are inadequate for protecting artificial intelligence systems, making it necessary to develop new, effective security strategies specifically designed to address AI-driven vulnerabilities.

---

2. Background: What is AI Security?

AI security refers to the protection of artificial intelligence systems from malicious attacks, unauthorized access, and unintended behavior. These systems are architectured to utilize machine learning models trained on large datasets to make predictions or decisions. The heavy reliance on data introduces significant risk into the entire system. If the data used to train or operate an AI system is manipulated, the system's outputs can be compromised.

Additionally, many AI models function as "black boxes," meaning their internal decision-making processes are difficult to interpret; their final output is released to the user without justification on how the model came to that decision. Due to this 'black box' nature, AI systems that have manipulated datasets can continuously produce incorrect or adverse outputs, which if gone unnoticed can compound to have tremendous negative impacts on society. The lack of transparency makes it challenging to detect and prevent attacks. As a result, AI systems present a broader and more complex attack surface compared to traditional software systems.

---

3. Key Security Threats in AI

3.1 Data Poisoning

Data Poisoning is a unique attack or vulnerability that has recently emerged with the development of artificial intelligence and machine learning (ML) systems and models. Resulting from AI's innate characteristics in which it relies on extremely large amounts of data, these datasets can be manipulated to affect the model's outcome. An attacker that is attempting to poison a dataset will purposefully and maliciously alter the dataset that is utilized to train an AI or ML model in an effort to achieve a specific outcome or output; the attacker aims to influence the behavior of a model to align with its conventionally negative prerogative or agenda.

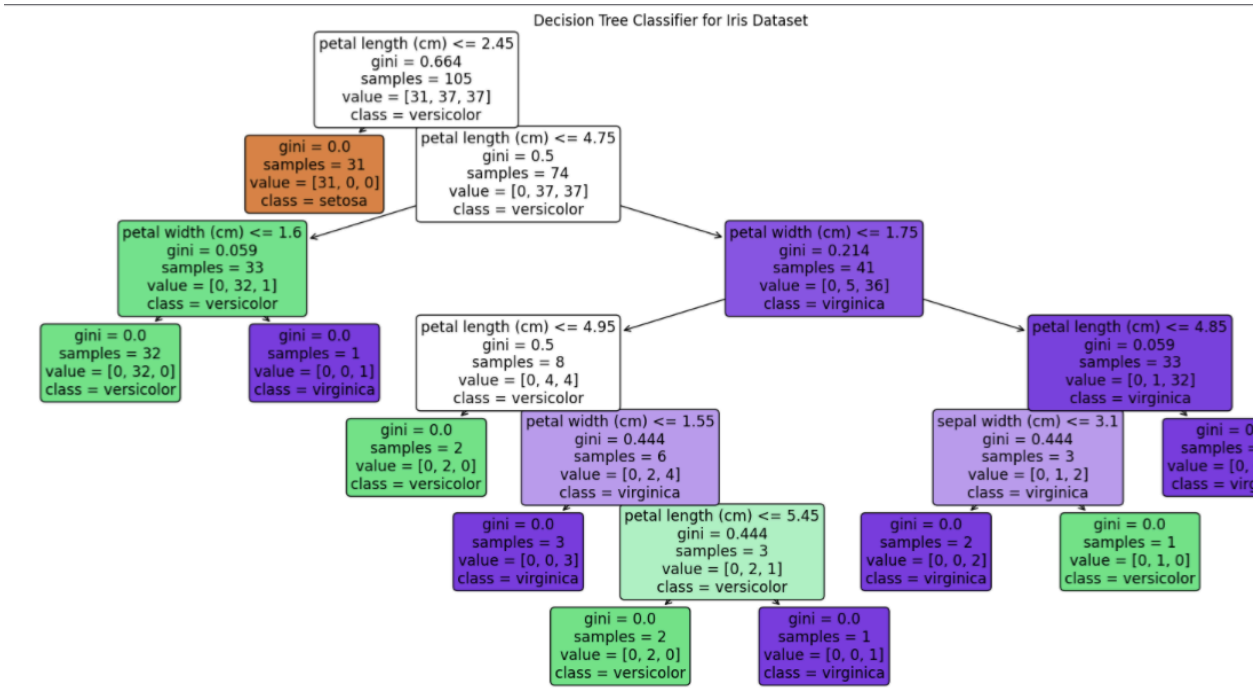
Artificial intelligence models are architectured to learn patterns from data and hence make predictions on those patterns. If that data is incorrect or misleading, the model will see a large, compounding degradation in accuracy and performance of outputs and prediction. This would be if you were taught incorrect information in law school and proceeded to become a lawyer and give incorrect legal advice to your clients.

When malicious attackers attempt to poison data they do so by adding incorrect, biased, or deceptive information into the datasets that the particular model they want to exploit are trained on. Throughout this process the model proceeds to make incorrect associations and assumptions which results in outputs that can be misleading, biased, or entirely wrong depending on the situation. This is dangerous, especially in today's environment for a few reasons. It is intrinsically dangerous as most models are "black box" models, including popular Large Language Models (LLM's). This is especially dangerous in today's world as many large LLM's such as ChatGPT, Gemini, or Claude have achieved such a reputation for themself in which some users wholeheartedly trust the model's output.

As researchers we often use visualization techniques, such as decision trees to verify the model's decisions and ensure it is coming to logical conclusions. However, as most popular AI models such as these aforementioned, wildly popular LLM's are "black box" models in order to protect their proprietary decision-making architecture, unique and sought after, we are unable to clearly verify the chain of decisions made to easily ensure models are not poisoned.

As pictured below *Figure 1* represents a decision tree classifier for the Iris dataset. Observed in the branches are specific, verifiable decisions that allow us to ensure the model is making correct associations and decisions. The tree allows us to actually visualize the entire process and hence trust that the data has not been tampered or poisoned.

Figure 1 represents the classification of irises.



If an attacker wanted to send malicious phishing emails or campaigns to an organization that utilizes an AI email filtering system they could manipulate the training data, as to classify malicious emails as benign. The overarching impact of data poisoning, because as previously mentioned the large majority of popularly used AI models are architectured in a way that makes it innately difficult to detect data poisoning. Once a model is subsequently trained on poisoned or compromised data, its outputs, although compromised, may appear legitimate. This classifies the attack in a unique category where it is borderline undetectable and persistent. Unlike many common attack vectors it can cause drastic societal impacts that may be untraceable and irreversible. If gone unchecked this can have tremendous, compounding effects on society that we may not even be able to anticipate the extremity of at this stage in AI development. This can directly undermine trust in AI systems and highlights a need for serious reconsideration of computer standards.

3.2 Adversarial Attacks

Adversarial attacks are another unique class of attacks that artificial intelligence systems face. To commit an adversarial attack an attacker intentionally crafts their input data to cause the model they are inputting data into to produce incorrect or deceptive outputs. Differing from data poisoning, adversarial attacks do not target the training phase, they instead occur during the model's operation. The changes to input data are noticed by the model and affect the model's decisions, hence triggering an incorrect or misleading output from the model. The attacker using this attack vector alters these inputs so they are imperceptible by humans.

---

4. Challenges in Securing AI Systems

The basis of modern cybersecurity frameworks originates from early computer information standards produced by the Department of Defense (DoD). The *Trusted Computer System Evaluation Criteria (TCSEC)*, known as the *Orange Book,* formalized system security evaluation methods and became the foundation of the Rainbow Series (Lipner, 2015). The Rainbow series then became the first major cybersecurity framework, which a majority of framework fundamentals are based on today. We still rely on the Rainbow series today, over 40 years later. "The Rainbow Series is still being used as a standard within some government contracts (Wilhelm, 2010)." When it comes to typical cybersecurity exploits, certain defense fundamentals from frameworks such as the Rainbow Series still suffice as exploits still follow similar tactics.

However, AI is a different domain and should be treated appropriately. Although AI is built on common foundations of computer science, the complexity of these models is drastically different from traditional systems we are familiar with. With this added complexity, we must exponentially increase the robustness and effectiveness of our artificial intelligence security and frameworks. We cannot continue to rely on tactics and processes drawn up 40 years ago. Although some of these processes that originated with early computer information standards, like the Rainbow Series are irreplaceable, the increased complexity involved with AI systems calls for not only a drastic rewrite, but a drastic redesign of frameworks. Due to the nature of vulnerabilities created by AI systems we require new standards to protect against exploits that the authors of the Rainbow Series simply could not have predicted.

Securing AI systems presents several challenges that go beyond traditional cybersecurity concerns. "The statistical, data-based nature of ML systems opens up new potential vectors for attacks against these systems' security, privacy, and safety, beyond the threats faced by traditional software systems (Vassilev, 2025)." NIST recognizes that there are technical challenges in securing AI that set it apart from the rest of cyberspace. Their nature of being data-based sets these systems apart from past information systems, with the AI Boom we see a new domain that needs to be treated differently. All systems can not be classified under the same

umbrella anymore. This idea presented by NIST affirms our aforementioned claims that AI systems are vulnerable to a host of new attack vectors due to their inherent nature of being data-driven systems.

Due to this nature they are vulnerable to a host of attacks unique to AI, including the previously mentioned adversarial attacks, data poisoning attacks, model theft, and model inversion. Beyond this, there are many zero day vulnerabilities that likely exist and attacks that will be relevant in the following years and decades that AI continues to develop. Although NIST recognizes that the thesis of this paper is true, they have done little to support it. They created the AI RMF 1.0; however, if they recognize that AI systems are inherently different from classic systems, they must also recognize that security and defense practices need to be inherently different.

There is a lack of standardized security practices specifically designed for AI. Many organizations rely on existing cybersecurity frameworks that do not account for AI-specific risks. There is also a real shortage of frameworks, each existing framework has issues. The European Union (EU) AI Act is tailored towards high-risk systems, leaving a large majority of developers feel like the act is not well suited for them (Dotan et al., 2024). According to Dotan 2024, this leads most developers to use either traditional frameworks or the NIST AI RMF 1.0, as the AI RMF is the only other existing major artificial intelligence framework. Although it addresses a broad range of AI risk dimensions, including privacy and governance, it places greater emphasis on high-level risk management processes than on detailed technical security guidance (OpenAi, 2026).

This speaks towards the challenges and shortfalls that AI developers are presented with when attempting to secure artificial intelligence. While it is absolutely necessary to focus on diminishing bias and ensuring privacy and fairness, there is a legitimate, crucial need for a document or platform that focuses on the unique technical challenges that come with AI. AI does also pose unique challenges regarding bias and fairness that we have not faced in prior systems, but that should not entice us to neglect AI security. Instead, we should strive to bind a group of professionals that are focused on the cutting-edge of AI security together to address these shortfalls and concerns. The rapid pace of AI development makes it difficult for security measures to keep up. New models and techniques are constantly being introduced, often without sufficient time for thorough security evaluation. Our frameworks and policy are already far behind technology. It is imperative that with the rate of change we are experiencing that we have an organization solely focused on AI security.

4.1 Limitations of Current Frameworks (NIST AI RMF)

The National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) represents a significant effort to address the risks associated with AI systems. It emphasizes principles such as trustworthiness, accountability, and risk management, providing organizations with guidance on how to approach AI governance.

However, despite its value, the framework has notable limitations when it comes to technical AI security. Its lack of detailed technical guidance limits the framework from being beneficial to a lot of companies. The framework focuses heavily on governance and risk management processes but provides limited direction on how to defend against specific threats such as adversarial attacks or data poisoning. While the framework does a tremendous job of framing governance and risk management, it is extremely lackluster in regards to security.

The framework does not fully address the rapidly evolving nature of AI threats. Attack techniques are constantly changing, and high-level guidelines may struggle to keep pace with emerging vulnerabilities. The National Institute of Standards and Technology (2023), says this in regards to AI security, “Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure. Guidelines in the NIST Cybersecurity Framework and Risk Management Framework are among those which are applicable here.” This is the extent of what the premier AI frameworks says in regards to security. It refers to frameworks that are still based upon the Rainbow Series, a set of publications that is greater than 40 years old; it does so in a manner that seems like an afterthought. The NIST AI RMF 1.0 doesn’t give any recommendations on how to tailor or implement the NIST Cybersecurity Framework specifically for AI systems. This gives conclusive evidence that developers have a lack of tools and resources at their disposal.

As opposed to the EU AI Act, the AI RMF is voluntary and not enforceable. This means that all of the guidelines are merely suggestions and organizations can pick and choose which pieces they follow if any at all. These conglomerated limitations demonstrate that even modern frameworks are not yet sufficient to fully secure AI systems, reinforcing the need for more specialized and technically focused security approaches that are effective and enforceable.

---

5. Proposed Solutions and Strategies

Addressing AI security challenges requires a combination of technical, organizational, and policy-based approaches. There needs to be baseline security requirements that are cleanly outlined in an enforceable policy-document. While it is incredibly beneficial to have optional guideline documents, AI is such a powerful tool that private sector organizations need to have

mandatory security requirements, so that there are repercussions for situations where an organization puts their bottomline above everything else.

It would be incredibly beneficial to have a framework or set of standards that focuses on the innate technical risks of AI models. It should focus on the technical aspects of items addressed in the AI RMF 1.0. Ensuring the integrity of training data is critical. Organizations should implement data validation processes, monitor datasets for anomalies, and restrict access to sensitive data sources. AI models should be protected through encryption, access controls, and secure deployment environments. Limiting access to models can reduce the risk of theft and unauthorized use.

There are a variety of methods where developers can implement adversarial training to make models more resilient against adversarial attacks. Since there are many ways to combat adversarial attacks it is difficult to know which combination of defenses you should use. Each defense mechanism improves robustness and accuracy by a few percent, so it is not enough to utilize one defense mechanism. Issues like this are where it would be incredibly helpful to have a compiled framework that provides well-researched security defenses and techniques for developers and AI executives to reference. According to , the best precaution to take against adversarial attacks is a multi-faceted approach. Developers should utilize defense methods, adversarial training, and intrusion detection in a cohesive manner to properly build robust AI systems.

Similar to typical security practices, simulated attacks should also be utilized. However, I feel that is much more crucial in this space. Due to the black box nature of most models, users cannot report vulnerabilities, so it is important to test for them.

5.1 Policy and Governance

As scholars, developers, politicians, and as a society it is our responsibility to create ethical, safe, and secure AI. While politicians and government agencies have put attention towards ethical AI in published frameworks, security has been overlooked. Hence safe has been neglected as well, as you can’t have safe AI without it being secure first. There is a potential for AI to be weaponized on a level we haven’t seen since nuclear war was created. It is absolutely pertinent that AI security be heavily considered before a major tragedy occurs. Politicians should work on creating a National AI Security strategy; it is imperative to notate that there is a significant difference between AI security and traditional cybersecurity. As academics and developers it is our duty to advocate for this, it is our duty to ensure we push the motivations to lawmakers. All organizations need to adopt clear, well-articulated policies for secure AI

development, deployment, and maturity. However, there needs to be diligent collaboration between industry leaders, academia, and government to establish a clear set of standards.

---

6. Future Outlook

Throughout the continued development of AI, as it becomes more complex and solidifies itself the challenges involved with securing AI will become more advanced. As AI becomes more intertwined with critical infrastructure services, the consequences of attacks will heighten. It is our opinion that this is when lawmakers will begin to take this risk more seriously. As support for AI Security will rise after these events, there will likely be a push for developing standardized frameworks specifically for security concerns, instead of broad AI governance frameworks. There will also likely be a need for a framework that considers security throughout the entire AI lifecycle. Depending on the severity of the sociopolitical environment at the time there may be a need for regulatory involvement as well. It is our mission to expedite this push. Instead of waiting for the consequences of inaction, we must take actions in accordance with our expectations of the future.

---

7. Conclusion

Artificial intelligence presents significant opportunities, but it also introduces new and complex security risks. Traditional cybersecurity approaches are not sufficient to address these challenges, as they were not designed with AI systems in mind; they were developed before AI existed. It is imperative that we evolve our security standards adjacently with our technological advancements. Although frameworks such as the NIST AI RMF 1.0 represent important progress, they do not properly address the technical vulnerabilities inherent in AI systems. As a result, there is a clear need for the development of specialized security strategies tailored to AI. Addressing these challenges will be essential to ensuring the safe, reliable, and secure deployment of artificial intelligence in the future.

References

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.15229

Lipner, S. B. (2015). The Birth and Death of the Orange Book. *IEEE Annals of the History of Computing*, *37*(2), 19–31. https://doi.org/10.1109/mahc.2015.27

National Institute of Standards and Technology. (2023, July 12). *AI Risk Management Framework*. NIST. https://www.nist.gov/itl/ai-risk-management-framework

OpenAi. (2026). ChatGPT 5.3 [Large Language Model]. https://chatgpt.com/

Scholz, M. (2026). *AI project in Google Colaboratory* [Computer code]. Google Colab.https://colab.research.google.com/drive/1UnXE8wCXAg6BPW8pkScOrYQKStMsSkY7

Shayea, G., Zabil, M., Habeeb, M., Khaleel, Y. & Albahri, A. (2025). Strategies for protection against adversarial attacks in AI models: An in-depth review. *Journal of Intelligent Systems*, *34*(1), 20240277. https://doi.org/10.1515/jisys-2024-0277

Vassilev, A. (2025). *Adversarial Machine Learning:* https://doi.org/10.6028/nist.ai.100-2e2025

Wilhelm, T. (2010). *CHAPTER 3 - hacking as a career* (T. Wilhelm, Ed.; pp. 43–99). Syngress. https://doi.org/10.1016/B978-1-59749-425-0.00007-5