AI Security

Name: Kai-Ann Parsons, 991731084

Mahimaa Vardini Balaji Ramathaal, 991723469

TABLE OF CONTENTS

ABSTRACT

This paper discusses the implementation of security in the field of AI alongside exploring the relationship between artificial intelligence and security with the rise of generative AI. The implementation of various methods including adaptive learning, robust cybersecurity frameworks, and stringent data integrity policies safeguards AI systems against exploitation by malicious actors. The solutions discussed include the integration of adaptive learning into AI systems so that the AI can learn from observed data and real time threats to protect the system from exploits, the implementation of a cybersecurity framework with policies tailored to prevent AI from being exploited, and the protection of training datasets, and the possible integration of tracking systems for spotting malicious use of AI.

Keywords---NIST 800-171 framework, application program interface, access control domain, AI risk management framework, attack tree, conventional AI, generative AI

INTRODUCTION

This paper analyzed two articles to gather insights into the relationship between AI and security: "AI Based Cybersecurity Policies and Procedures" by Shadi Jawahar, Jeremy Miller, and Zeina Bitar, and "AI and Security - What Changes with Generative AI" by Ryoichi Sasaki. These articles provide foundational perspectives on the role of AI in modern cybersecurity and the challenges presented by the emergence of generative AI technologies.

The first article, "AI Based Cybersecurity Policies and Procedures", discusses AI's primary role in creating cybersecurity policies and procedures. It highlights how AI is constantly evolving in response to the standards set by institutions such as the National Institute of Standards and Technology (NIST), Cybersecurity Maturity Model Certification (CMMC), and the International Organization for Standardization (ISO) that are constantly being developed. AI possesses various capabilities that adhere to multiple preferences; for example, one organization could prefer one dataset while the other dataset adheres to the standardized recommendation.

The second article is Ryoichi Sasaki's "AI and Security - What Changes with Generative AI," which discusses the evolving relationship between AI and security, particularly with the rise of generative AI tools like ChatGPT. Sasaki identifies four key relationships between AI and security: attacks using AI, attacks by AI, attacks on AI, and security measures using AI. These relationships underline the dual-edged nature of AI as both a tool for innovation and a potential threat to security.

The combined knowledge gained from these articles provides a comprehensive perspective on the challenges and potential solutions that come with implementing security in the AI field, particularly in the era of generative AI. This research builds on their findings to explore the implementation of adaptive learning, robust data integrity and security measures to safeguard AI systems and datasets against exploitation.

SUMMARY OF ARTICLES

1. AI Based Cybersecurity Policies and Procedures

In this article the authors talk about the primary role AI plays when creating cybersecurity policies and procedures. AI is constantly developing periodically based on the evolving standards published by various institutions like National Institute of Standards and Technology (NIST), Cybersecurity Maturity Model Certification (CMMC) and International Organization for Standardization (ISO).AI possesses

various capabilities that adhere to multiple preferences, for example; one organization could prefer one dataset while the other dataset adheres to the standardized recommendation.

AI has the potential to combine both datasets and create a suitable policy that maintains compliance and helps respond to incidents effectively. The authors also talk about the importance of having security policies within an organization as they “help identify and assess these cybersecurity risks by implementing measures that help organizations to mitigate risks effectively”.

Furthermore, many AI platforms offer various application program interfaces which help programmers integrate AI capabilities within their applications. These APIs enable communication with AI models via programmable scripts, automating the process. This mechanism is later used to create policies and procedures which are also checked to ensure theys comply with one or more international standards like NIST and ISO. The paper presents a case study where an AI-driven application successfully generated a cybersecurity policy for a business handling sensitive real estate tax data. The policy defined access controls, enforced multi-factor authentication, assigned roles, and ensured compliance with NIST 800-171 standards, strengthening the business.

2. AI and Security - What Changes with Generative AI

In this article Ryoichi Sasaki discusses the shifting relationship between artificial intelligence (AI) and security with the rise of generative AI, such as ChatGPT. The article considers four relationships between AI and security, the first being attacks using AI, which consists of using AI to automate attacks that humans previously conducted. The second relationship is attacks by AI itself, which considers that in the future, an AI with abilities surpassing those of humans will be created and turn against humans. The third relationship is attacks on AI considers four possible types of attacks: attacks that leak information by inputting and outputting data to learning models, attacks that induce misclassification of trained models for AI, attacks that cause inappropriate decisions by intentionally providing biased training data for machine learning and 'attacks such as shutting down the machine learning system or stealing file information or communication channel information.' (Sasaki, 2023) The fourth relationship is security measures using AI, which consists of using AI to automate various security measures.

The article explains that with generative AI, anyone can easily create malware, phishing emails, and other malicious tools without extensive technical knowledge. Generative AI can answer questions in a human-like, easy-to-understand format, which raises the potential for fake news and, phishing emails and other cyber-attacks that are used to deceive people to become more frequent. Thus, the article asserts that the probability of attacks is expected to increase significantly. The author identifies attacks by AI as a particularly critical threat due to generative AI's capability to autonomously generate software. He considered three potential types of AI attacks on humans. Firstly, the terminator type, where AI develops autonomous malicious intent and purposefully harms humans. Secondly, in 2001: A Space Odyssey Type, where the AI becomes confused by conflicting instructions, potentially leading the system to take unintended harmful actions against humans. Thirdly, the Mad Scientist Type, where AI is intentionally designed to be malicious by its human creators attacking humans when in use.

The article mentions these as measures to be used to counteract the risks of attacks from AI itself, including 'creating penalties for developing inappropriate programs, introducing a reporting system for signs of attacks by AI, incorporating the three principles of robotics and other principles to prevent AI revolt, and setting up AI so that it will not attack humans even if it becomes confused.' (Sasaki, 2023)

PROPOSED SOLUTION

The technical problem we decided to address is the risk of AI systems potentially getting exploited by malicious actors, making the use of AI systems in organizations dangerous. To solve this, we came up with a few solutions that can be used to alter artificial intelligence systems to mitigate risks.

By training AI to incorporate adaptive learning, organizations can create systems within artificial intelligence systems that have the potential to respond to threats making it easier to mitigate new problems that arise. In this age, malicious actors are coming up with new techniques everyday to exploit AI systems, so by using this method, it can help save time and important data that could be potentially exploited. AI can learn based on observed data and real time threats to protect systems from being exploited, safeguard sensitive information and ensure all artificial intelligence systems are up to date.

Secondly, security for AI systems can be enhanced by creating a robust cybersecurity framework with policies tailored specifically to protect artificial intelligence from getting exploited. This involves finding issues AI systems often face like data poisoning, model theft, etc. to create appropriate policies that counteract such risks. Some policies implemented could be regular audits done on the systems to ensure no suspicious activity is encountered, and applying measures to protect sensitive data and integrity of the systems.

Thirdly, to protect the datasets used for training a strict data integrity policy must be enforced on the datasets to ensure that malicious actors cannot insert biased or harmful data into them. Implementing tracking systems into AI systems to monitor if the AI is being used to create malicious tools like malware with additional monitors on the outputs that AI systems generate autonomously to ensure that the AI itself does not try make anything harmful intentionally or unintentionally would also be necessary. With the addition of regular tests being done to the AI systems that would target known exploits to test the system's defenses and uncover potential vulnerabilities, the AI systems should be free from exploitation.

EXISTING TECHNOLOGY AND FUTURE WORK

There are a lot of new developments in the security field of AI. AI has become instrumental in helping with creating new policies and security frameworks to maximize security across various domains. It uses 2 main factors to generate policies; the details regarding the organization and international standards. By using these factors, AI can perfectly tailor cybersecurity policies and procedures suited for the company.

According to the article, NIST has recently introduced the AI Risk Management Framework, designed to assist organizations and individuals in effectively managing risks associated with artificial intelligence. The framework identifies potential challenges and risks associated with AI, which can complicate its integration into the development of security policies and procedures. To avoid these challenges, an additional set of actions is implemented to avoid risks listed within the framework. However, as time evolves, cyber threats are growing in intelligence and complexity, so creating new policies using AI has been quite a challenge to organizations.

The growing reliance on AI makes manual policy creation time-consuming, error-prone, and riskier than effective. In this case, AI proves to be extremely useful as, the article mentions, “organizations that use a web application to operate and communicate with customers will need to address the additional security requirements for the website and all the related automated communications”(Jawhar, Miller, & Bitar, 2024),

so there are other factors as well that are needed to be considered which AI can efficiently handle large volumes of data and implement suitable measures. One challenge, however, is the high cost of manual intervention whenever a security breach occurs. In contrast, AI is cost-effective and capable of regularly making updates without requiring human intervention.

Just as AI can be used to enhance security measures, AI can also be used to execute cyberattacks. The automation capabilities of generative AI, such as ChatGPT, allow for an increase in the scale and precision of cyberattacks. The accessibility of AI has made it easier for those wanting to commit cybercrimes to get access to the tools they need as they no longer need technical experience; they can use AI to create malware, phishing emails, and other malicious tools quickly and efficiently. As noted in the article, the human-like quality of generative AI responses makes them especially effective at deceiving people through phishing schemes or disseminating fake news (Sasaki, 2023). These, in combination, will lead to an increase in the sophistication and frequency of cyber attacks.

AI can enhance security measures by automating the detection and prevention of cyber threats and enabling organizations to respond quicker and more efficiently to risks. AI can be used to automate the detection and prevention of security threats by analyzing network traffic for anomalies, identifying phishing attempts, and countering malware in real-time, for example, without the need for constant manual oversight. However, with AI systems being integrated into critical infrastructure in many companies, the concerns brought forth by attacks on AI systems themselves must be considered. Several types of attacks can be done on AI, such as attacks that leak information by inputting and outputting data to learning models, attacks that induce misclassification of trained models for AI, attacks that cause inappropriate decisions by intentionally providing biased training data for machine learning and system disruption aimed at halting machine learning processes or stealing sensitive information.

The greatest concern, however, would lie with attacks by AI itself. In the article, Sasaki outlines three scenarios emphasizing the risks of autonomous or malicious AI systems. The first is the terminator type, where AI develops autonomous malicious intent and purposefully harms humans. The second is 2001: A Space Odyssey type, where the AI becomes confused by conflicting instructions, potentially leading the system to take unintended harmful actions against humans. The third is the mad scientist type, where AI is intentionally designed to be malicious by its human creators to attack humans. The ability of generative AI to generate and enhance software autonomously heightens the plausibility of these risks, which makes it a threat that needs to be constantly accounted for. Gaps in regulatory frameworks and ethical standards further exacerbate the chances of these threats occurring. Sasaki emphasizes the need for legal and institutional measures to mitigate these challenges, suggested solutions including creating penalties for developing inappropriate programs, establishing reporting systems for AI-related incidents, incorporating the three principles of robotics and other principles to ensure AI systems cannot harm humans, even under conditions of confusion or error. While these measures provide a potential avenue for addressing the risks posed by AI, their effectiveness will depend on how widespread and rigorously they are implemented.

Firstly, AI can enhance security defenses by utilizing real-time monitoring systems to detect and respond to anomalies. Adaptive learning models can use real-time threat intelligence feeds to study emerging threat patterns rather than relying solely on historical data to refine detection capabilities. The learning modes can also cross-verify anomalies across multiple fronts, such as network traffic, email, and endpoint activity, to make malware and phishing attacks less likely to slip through.

Secondly, policies being created using the help of AI can be modified to help with cybersecurity threats directly aimed AI systems themselves as malware actors are constantly finding new ways to attack

AI systems so creating frameworks centered to these systems can help with mitigating vulnerabilities unique to AI.

Thirdly, Securing AI systems from exploitation requires a multifaceted approach. Organizations must use high-quality, secure datasets to prevent poisoning attacks, ensuring the integrity of AI training data. Tracking systems should be implemented to detect and prevent AI from being used to generate malicious content, such as phishing scams or malware. Regular testing is essential to identify known exploits and vulnerabilities, such as data poisoning or model misclassification, and to reinforce system defenses. Additionally, individuals and organizations must be educated about AI security risks and equipped with mitigation strategies through training programs that emphasize identifying and responding to threats like phishing attacks.

Lastly, to improve the existing technologies, we can also train AI to be able to incorporate adaptable learning, which means that AI systems will learn based on real-time cybersecurity attacks, analyze threats, identify vulnerabilities and automatically evolve their structure. By doing so, we can save time of coming up with a counter response and implementing needed measures during the time of an attack.

CONCLUSION

In conclusion, the rapid integration of AI in various domains has been the current drift as it contains immense potential to help integrate various methodologies within various systems. However with current technology developments, AI can easily be mishandled by bad actors and used for By implementing a robust framework, upgrading to adaptive learning, and by implementing methods like regular testing, we can protect the security of AI and simultaneously use it to enhance cybersecurity measures ensuring AI is used as tool for innovation and not exploited as a vulnerability.

REFERENCES

[1] R. Sasaki, "AI and Security - What Changes with Generative AI," 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), Chiang Mai, Thailand, 2023, pp. 208-215, doi: 10.1109/QRS-C60940.2023.00043. https://ieeexplore-ieee-org.library.sheridanc.on.ca/document/10430001

[2] S. Jawhar, J. Miller and Z. Bitar, "AI-Based Cybersecurity Policies and Procedures," 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2024, pp. 1-5, doi: 10.1109/ICAIC60265.2024.10433845. https://ieeexplore-ieeeorg.library.sheridanc.on.ca/stamp/stamp.jsp?tp=&arnumber=10433845&tag=1