Assignment: Research Paper on Artificial Intelligence in Cybersecurity

Michael Twining

Department of Criminal Justice Studies, Utica University

[REDACTED]

[REDACTED]

[REDACTED]

Abstract

Artificial Intelligence (AI) has become a cornerstone of modern cybersecurity, offering unparalleled capabilities for detecting threats, automating responses, and predicting vulnerabilities. However, vulnerabilities in large language models (LLMs) and machine learning (ML) algorithms expose organizations to significant risks. This paper examines the dual nature of AI in cybersecurity, highlighting both its potential to enhance defense mechanisms and the risks it poses when exploited by malicious actors. Through an analysis of recent case studies, emerging attack vectors, and industry frameworks, critical vulnerabilities are explored such as prompt injection, data poisoning, and model manipulation. The research emphasizes the systemic risks associated with AI deployment, including potential societal impacts and ethical considerations. By drawing insights from the National Institute of Standards and Technology (NIST) AI Risk Management Framework, Open Worldwide Application Security Project's (OWASP) Top 10 for LLMs, and other industry standards, this study proposes a comprehensive approach to secure AI adoption. Lastly, the paper outlines robust mitigation strategies and emphasizes the need for a balanced, risk-aware approach to integrating AI within cybersecurity ecosystems, ensuring resilience against evolving threats while maximizing the benefits of this powerful technology.

Introduction

Cybersecurity is a crucial element to any entity in the public and private sectors. Without it, threat actors simply run amuck seemingly unnoticed on information systems, stealing sensitive data and leaving potentially devastating impacts on those entities. According to the Federal Bureau of Investigation (2024), internet crime in 2023 resulted in $12.5 billion in losses. Among the top five categories of internet crime were phishing, personal data breach, non-payment schemes, extortion schemes, and tech support fraud (Federal Bureau of Investigation, 2024). In fact, among the top five sectors impacted in internet crime were healthcare and public health, critical manufacturing, government facilities, information technology, and financial services (Federal Bureau of Investigation, 2024). Even with the best cybersecurity practices, threat actors can still find ways to exploit vulnerabilities. In some parts, this can be due to large amounts of data needing to be sifted through by a human, who might not be able to correlate data as efficiently as a machine. The use of AI enables cybersecurity defenders to assist in this correlation analysis to call out potentially true positives, filtering out the noise that false positives or false negatives may cause. While AI offers powerful tools for enhancing cyber defense capabilities, the vulnerabilities in LLMs and ML algorithms pose significant risks. Effective implementation of AI in cybersecurity requires a thorough understanding of these systemic risks and the development of robust mitigation strategies. This research will investigate both the advantages and disadvantages of AI in cybersecurity, concluding with a discussion of essential risk mitigation and ethical considerations.

An Overview of How AI Works

AI has specific functions aimed at achieving specific tasks. Some of the fundamental components are aimed to sense, reason, and act. Sensing is when the AI obtains input from a

sensor or data feed. From there, it uses algorithms to process the input to make decisions or predictions based on patterns and classifications of the input compared to the trained data it has categorized in its own data set. This is where machine learning occurs. The AI then acts once it has provided reasoning and reached a conclusion (Hammond, 2015). Much of the functionality and features of AI occurs in what it senses in data and how it reasons with the input, making AI appear to have learning capabilities of its own, also known as deep learning. There is also natural language processing that enables the AI to understand and generate a response in a human language, rather than the technical processing it uses to interpret a response (Hammond, 2015). This is particularly useful in chatbots. Lastly, there is the ability to perform predictive analysis, where the AI has reasoned based on historical data to predict something before it happens (Hammond, 2015). It is for this reason that AI is very useful in cybersecurity, as it may provide a correlation of multiple events as a specific cyber attack in a short amount of time compared to human analysis.

The Role of AI in Cybersecurity

AI has become an essential tool in cybersecurity, offering advanced capabilities for threat detections, predictive analytics, and automated responses. As threats grow in complexity and frequency, public and private sector entities will need to bolster defenses with AI to optimize resources and enhance overall resilience against more sophisticated attacks. SentinelOne (2024) reports that threat detection systems using AI increase accuracy rates between 80 to 92 percent (2024). This is due to the AI-fed threat intelligence used to identify markers of malicious activity. ML algorithms can sense, or analyze, and reason from extremely large datasets, which can help to detect phishing, malware, and various other indicators of compromise (IoCs).

In addition to detecting threats, AI enables predictive analytics by correlating data to identify patterns of behavior indicative of potential risks. AI prioritizes resources by identifying high-risk patterns and filtering out less severe issues, allowing security teams to focus on critical threats. Organizations using AI extensively for prevention reduced the average cost of a data breach by 45 percent, from $4.88 million to $2.22 million (SentinelOne, 2024). In some cases, the ability to contain incidents to minimize impact can be the largest factor in the cost of a cyber incident. Predictive analytics accelerates incident containment. This is crucial as it takes a security team an average of 277 days to identify and contain a breach, and 328 days for any breaches that involve lost or stolen credentials (SentinelOne, 2024). This overwhelming volume of data makes it nearly impossible for human teams to respond to breaches effectively without AI's ability to process information at scale and in real time.

As mentioned previously, AI is a tool with many uses in cybersecurity. In 2023, a vulnerability was reported every 17 minutes according to SentinelOne's (2024) analysis of Common Vulnerabilities and Exposures (CVEs) from the National Vulnerability Database (NVD). This creates an immense amount of data that it seems only a machine can effectively interpret and apply to monitoring networks and endpoints. A notable example is Cisco SecureX, which leverages AI to detect anomalies across networks and automate responses such as quarantining malicious files or shutting down compromised endpoints (Perception Point, n.d.). This type of automated response is like having an additional teammate on the cybersecurity team. For instance, an AI model might identify multiple failed login attempts from a specific IP address and automatically blacklist it or quarantine an endpoint when ransomware-related IoCs are detected. By automating routine security tasks and incident responses, AI reduces reaction times while minimizing human error.

Among the uses mentioned previously, there are traditional uses of AI in cybersecurity as well. For example, with endpoint protection, Cylance uses machine learning models to proactively block malicious files without relying on traditional signatures (Kaspersky ML Research Team, 2021). Another use is network monitoring, where SentinelOne integrates endpoint protection with real-time threat detection to provide comprehensive coverage against advanced persistent threats (SentinelOne, 2024). AI can help strengthen user authentication as well, through tools used for facial recognition and biometrics, especially when there are indicators of fraudulent login attempts.

The benefits of AI in cybersecurity tasks are immense as threats are only growing over time. The ability to filter out noise in alerts is a major benefit. Organizations using extensive security AI reported a significant reduction in false positives compared to those relying on manual methods (OWASP LLM Project Admin., 2024). As mentioned previously on detection and response, AI can significantly reduce the amount of time to detect and respond to incidents, limiting the total impact of an incident or event. It also can act as a sentry, where a human may be off from work, the AI can continue to monitor, detect and respond to threats. The total cost savings from an event are noteworthy as well. The ability to reduce the annual financial impact from $4.88 million by $2.22 million cuts the financial impacts nearly in half (SentinelOne, 2024).

AI has revolutionized cybersecurity by enhancing threat detection capabilities, automating responses, enabling predictive analytics, and improving scalability. Its ability to process vast amounts of data in real time allows organizations to stay ahead of evolving cyber threats while optimizing resource allocation. This can aid in reducing the impact of attacks by groups like LOCKBIT, BlackCat, Akira, Royal, and Black Basta, who have been designated as

the most used ransomware variants affecting critical infrastructure in 2023 by the Federal Bureau of Investigation (2024). By leveraging tools like endpoint protection platforms, network monitoring systems, and advanced authentication mechanisms, organizations can strengthen their defenses against increasingly sophisticated attacks. The statistics underscore the transformative impact of AI: faster detection times, reduced breach costs, and improved scalability make it an essential component of modern cybersecurity frameworks. As adoption continues to grow, the integration of AI into cybersecurity will remain critical for safeguarding sensitive data and critical infrastructure in an ever-changing threat landscape.

Vulnerabilities Within AI

Artificial Intelligence (AI) has become a cornerstone of modern cybersecurity, offering advanced capabilities for detecting threats, automating responses, and predicting vulnerabilities. However, vulnerabilities in LLMs and ML algorithms expose organizations to significant risks. These vulnerabilities, ranging from adversarial attacks to systemic societal impacts, can be exploited by malicious actors, undermining the very systems AI is designed to protect. Figure 1 below illustrates just how big the attack surface is which exists within AI. The next few sections will break down these vulnerabilities, analyze specifically how threat actors exploit the ML algorithms and LLMs through case studies, and later discuss some calls to action in mitigating these threats.

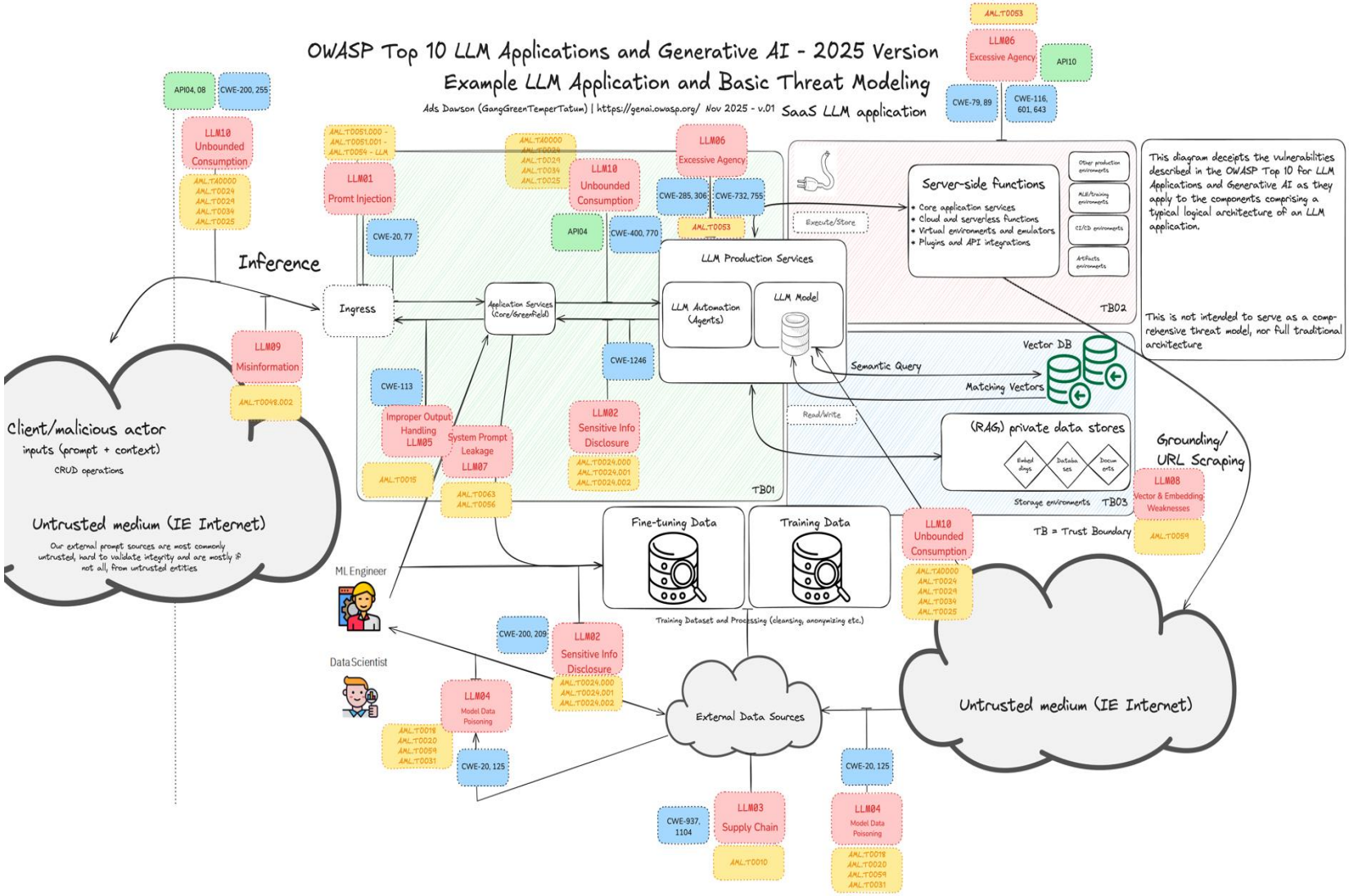

Figure 1. OWASP Top 10 Example LLM Application and Basic Threat Modeling. Adapted from *OWASP LLM Project*. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/. Copyright 2024.

Prompt Injection

Prompts are the inputs from a user to the AI. For example, if someone uses a platform like ChatGPT, they ask the AI a question, or in this case provides a prompt for the AI to respond to. Through prompt injections, the user is attempting to bypass any safety protocols that may prevent an AI from providing a response to the user (OWASP LLM Project, 2024). In essence, they are trying to trick the AI into providing a response through a cleverly crafted prompt. With direct prompt injection, the user's prompt is directly using the prompt to alter the behavior of the AI. Indirect prompt injection is where the AI accepts input from a source, a file upload for example, which may alter the behavior of the AI (OWASP LLM Project, 2024). Embrace the Red (2023) was able to successfully exploit an indirect prompt injection vulnerability within ChatGPT through an open-source plugin used by the AI. The team created a malicious website that ChatGPT used because it was prompted to do so. From there, the malicious website had an indirect prompt write the chat history of the user from ChatGPT to the URL, which could be exfiltrated later (Embrace the Red, 2023). This places the user at potential risk of PII exposure, as well as total invasion of privacy of any chat that has been recorded by the AI. Another example builds on the classic Morris Worm, which was an internet virus capable of self-propagation back in the late 1980s. Morris II is a zero-click worm that is designed to self-propagate within an AI ecosystem (Cohen et al., 2024). This was used to send a malicious email which contained the self-replicating prompt in the target system utilizing the AI (Cohen et al., 2024). In context this AI is used to provide responses to emails as a generative AI (GenAI). As the AI replies to the email, the prompt changes the behavior of the assistant, and the worm part of the email that has the malicious prompt can successfully extract information stored in the AI, potentially leaking sensitive information (Cohen et al., 2024).

Prompt injections pose a significant threat to AI security, particularly in cybersecurity. These case studies suggest human behavior should be required when running a prompt, using CAPTCHA for example, which would limit the impact of self-propagating programs and malicious sources from exploiting vulnerabilities in AI. Prompt injection exploits where self-propagating programs worm the system shows the need for segmentation between prompt data and source materials and the AI itself (OWASP LLM Project, 2024).

Sensitive Information Disclosure

As public and private sectors implement AI in various applications, there is an increased risk of introducing sensitive information into the LLMs that store the information to improve machine learning and future responses. Consider the scenario where an organization is using AI to analyze expense reports in Excel with an embedded AI application attached to the program. It is possible that information is stored into the LLM database outside of the entity, leaving sensitive information vulnerable to leakage. In the previous exploits through prompt injections, it is clear why this can be cause for concern, as this information can be exfiltrated by threat actors. From a traditional cybersecurity threat perspective, insider risks also exist. This is where an employee may knowingly or unknowingly contribute to the compromise of a system. When this occurs, the keys to the kingdom are essentially forfeited, especially if the victim has administrator rights to the target, in this case the AI components, such as the training data. The Microsoft AI Red Team (2020) discovered this capability to exfiltrate information through an exercise with some knowledge of how a specific Microsoft Azure Service environment was established. The team started with reconnaissance, which is the first step in the cyber kill chain. The cyber kill chain, developed by Lockheed Martin, outlines the stages of a cyberattack, from initial reconnaissance to data exfiltration, providing a framework for understanding and disrupting threat actor activities.

Through their research, The Microsoft AI Red Team was able to use a valid account to gain initial access to a network. From there, they discovered the targeted ML model along with the training data, and exfiltrated it remotely. Afterwards they crafted a way to evade detection within the training data which they replaced, used an exposed API to access the model, and submitted adversarial examples into the production environment (Microsoft AI Red Team, 2020). This enabled them to evade future detection and maintain some level of persistence to change the behavior of the AI.

Intended and unintended disclosure of information can potentially increase the level of risk a public or private sector entity is exposed to without them being fully aware. While encryption is an effective mitigation strategy against this type of attack from an insider threat, it may not be effective if they are also able to decrypt the information. Segmentation of the ML model and training data could also assist in limiting the impact of the compromise (OWASP LLM Project, 2024). In the example above, the training data and ML model should not have been stored in the same environment, as this made it too easy to manipulate the complete AI ecosystem.

Improper Output Handling

Improper output handling occurs when there is insufficient validation, sanitization, and handling of the output in AI before it passes into its other components and systems (OWASP LLM Project, 2024). This is similar to prompt injections, whereas prompt injections may be a vector of attack to exploit additional functionality of the AI. Imagine that a threat actor can utilize a chatbot service and use prompt injections to exfiltrate sensitive customer information stored in a database that feeds the LLM information to formulate responses. This was a proven concept in the example for indirect prompt injections by Embrace the Red team. Another example is exploiting a web app that uses AI to generate content from user prompts without sanitation, which could cause the AI

to return unsanitized JavaScript payloads, leading to an XSS attack when the victim opens the content on their browser (OWASP LLM Project, 2024). Another possible scenario is the ability of an attacker to use SQL injection to craft an SQL prompt to delete database tables, due to the prompts not being sanitized (OWASP LLM Project, 2024).

There are ways to help mitigate the threat of improper output handling. According to the OWASP LLM Project (2024), developers could implement additional authentication requirements or adopt a zero-trust posture with AI plugins, establish some form of validation and sanitization process before the prompt is read by the AI, and fine-tune AI models to be aware of abuse or malicious activity that creates unusual patterns of behavior in the AI itself.

Data and Model Poisoning

Data poisoning is when the data used for training, fine tuning, or embedding data with the AI is abused to introduce vulnerabilities, backdoors, or biases (OWASP LLM Project, 2024). The examples above have shown how threat actors access training data and other subcomponents of AI, making it an easy task to simply change the data. One example is split-view data poisoning. During this crafty tactic, the attacker hosts a domain that is expired by purchasing it after conducting due diligence to determine if it is known for being used in public training data for AI. Then they rebuild the domain to contain malicious content that the LLM will read from, as shown in Figure 2 with cat memes. The LLM now has thousands of cat memes that it is trained on, impacting the data the AI has been trained on (GangGreenTemperTatum, 2023b). If the domain was used for threat intelligence in a cybersecurity platform, it has now been tainted with data on cat memes and is not performing as expected.

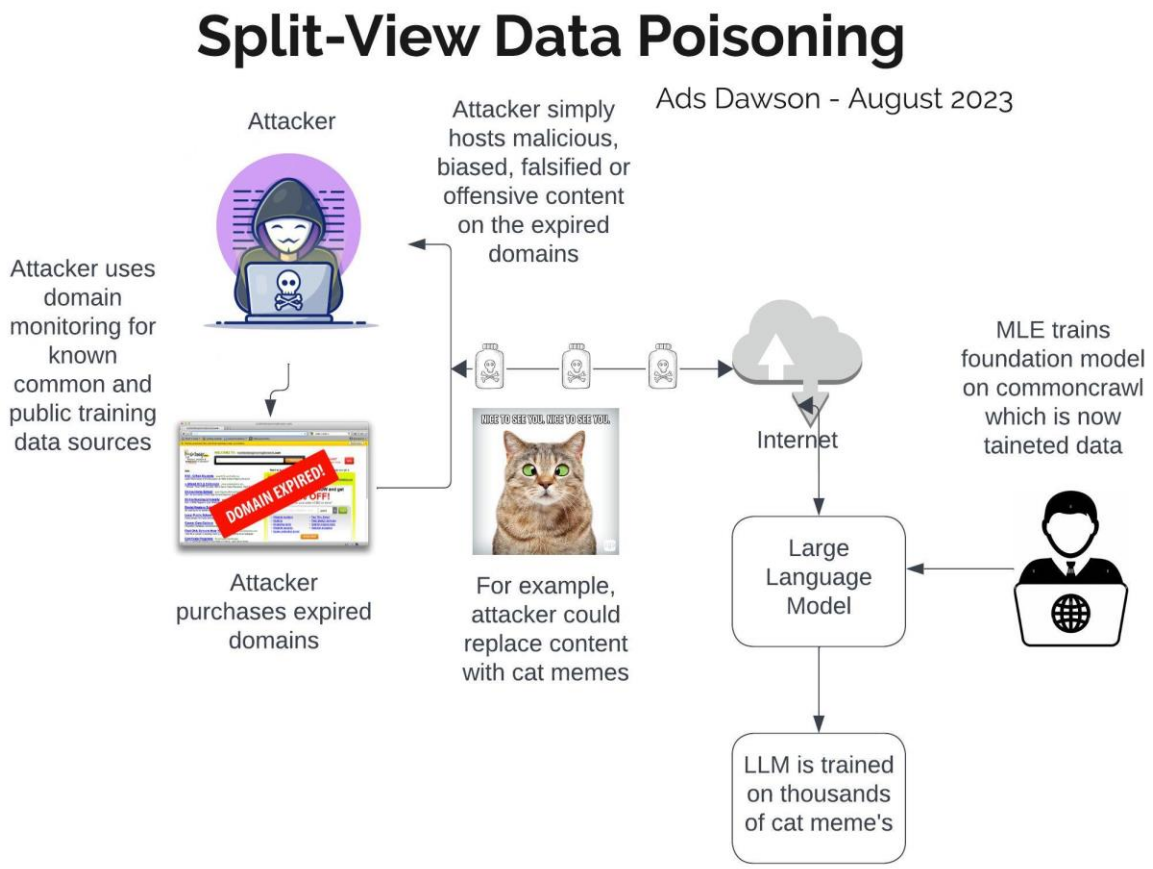

Figure 2. Split-View Data Poisoning. Adapted from *GangGreenTemperTatum*. https://github.com/GangGreenTemperTatum/speaking/blob/main/dc604/hacker-summer-camp-23/Ads%20_%20Poisoning%20Web%20Training%20Datasets%20_%20Flow%20Diagram%20-%20Exploit%201%20Split-View%20Data%20Poisoning.jpeg. Copyright August 29, 2023.

Another tactic, frontrunning data poisoning, is similar, however, instead of the threat actor buying an expired domain that is used as a public training data source for AI, the attacker reverse engineers the training data source itself. From there, they inject malicious content, misinformation, offensive content into the training data, tainting the LLM used by the AI, as shown in Figure 3 (GangGreenTemperTatum, 2023a).

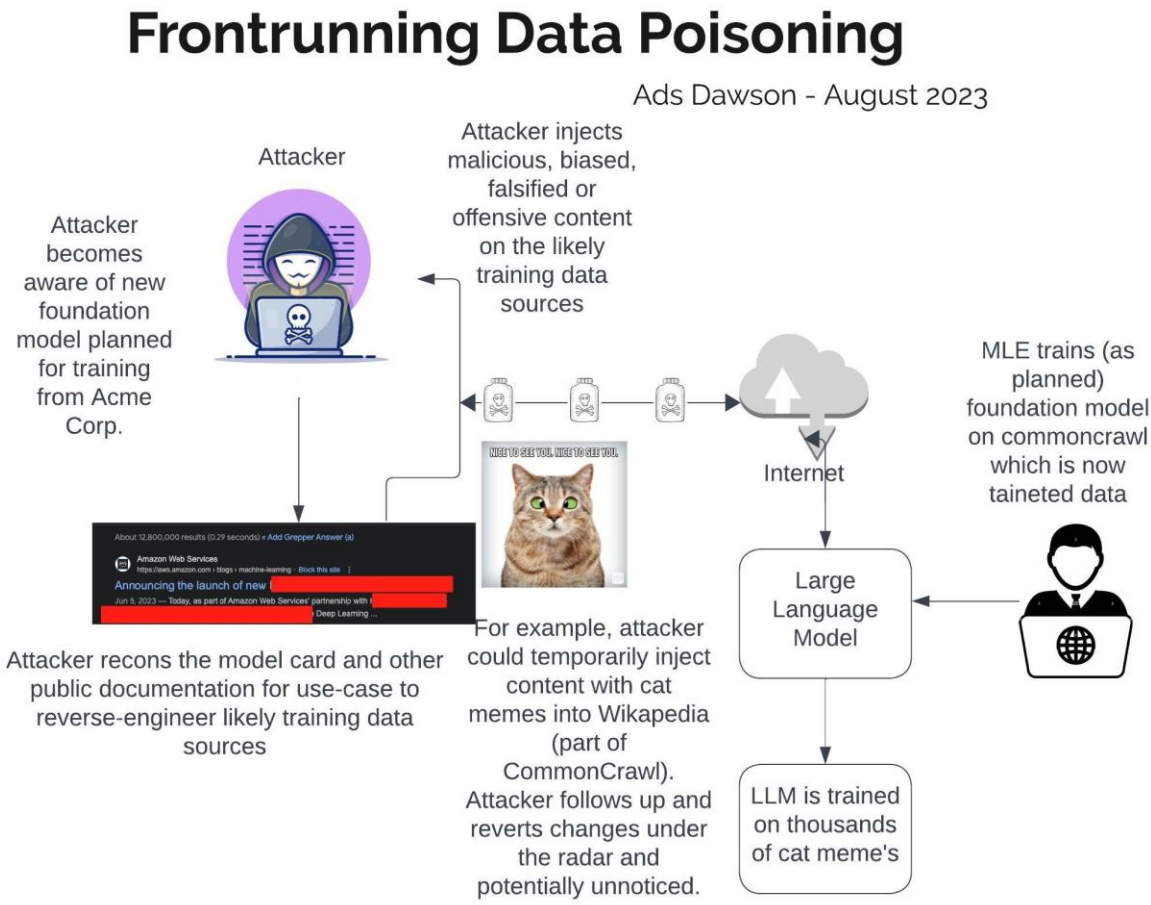

Figure 3. Frontrunning Data Poisoning. Adapted from *GangGreenTemperTatum*. https://github.com/GangGreenTemperTatum/speaking/blob/main/dc604/hacker-summer-camp-23/Ads%20_%20Poisoning%20Web%20Training%20Datasets%20_%20Flow%20Diagram%20-%20Exploit%202%20Frontrunning%20Data%20Poisoning.jpeg. Copyright August 29, 2023.

Where would the threat actor get information about what training sources were used in an AI model? Figures 2 and 3 reference Common Crawl, which is a public database with files on billions

of webpages and content. In fact, in January of 2025alone, the archive contained three billion pages (Common Crawl, 2025). This information is publicly available often on the AI provider's page.

Providers of AI have too much public transparency that makes the LLMs vulnerable to this type of attack. ChatGPT, for example, utilizes Microsoft Azure services to store training data, as advertised on their website. While they do not say what they use for training data, a threat actor could find exploits within the Microsoft Azure environment, possibly through cloud misconfiguration settings or access through a successful phishing attempt. This provides them with full access if it is the right account with the right privileges. In 2020, a researcher was able to gain access to Clearview AI's private code repository through a misconfigured cloud server, which allowed for arbitrary users to register a valid account (Researchers at spiderSilk, 2020). Once they created the account, users were able to access an S3 bucket hosted on AWS cloud, which contained all of the information used for this facial recognition tool used by law enforcement. The attacker was able to download all the training data and information about software, models, and capabilities from the source code, including the ability to decompile the application binaries (Researchers at spiderSilk, 2020). As a result, the researchers could have altered the training data or simply deleted the databases, or worse yet, found a way to establish persistence and listening abilities to the AI through a back door.

Misinformation

Misinformation is not solely an issue in the media, but also in technology. Specifically, AI knows what it is trained to know, so it would not be able to distinguish misinformation from correct information. Misinformation in AI can occur from a breach of the LLM where data has been modified, through hallucinations where the AI provides an unsupported assumption to provide a nice-sounding response to the end-user, or from bias present in the training data (OWASP LLM

Project, 2024). As dependence on AI increases, this is a major concern, as misinformation can spread exponentially. This can occur in LLMs that have been breached where the attacker has altered the training data to disrupt the behavior of the AI. The Kaspersky ML Research Team (2021) was able to exploit this vulnerability to confuse an AI used for anti-malware. The team researched techniques used by others to attack AI which was used and trained on malware detectors, along with identifying on the antimalware's website that the program was ML-based. After using the antimalware software, they began to understand how it worked, learning that it extracts a local system's features and sends them to a cloud-based AI, where it scans malware detectors and classifies them. The team used the dataset of malware and clean files to scan in the target AI solution and change the labels of the samples. Once they saw how the AI acted, they learned the AI collected PE header features of the executable file, as well as file strings and section features to provide a number classification of the file. They created an adversarial algorithm that could trick the AI by changing features of the file and maintaining the payload to confuse and bypass the detection of the malware (Kaspersky ML Research Team, 2021). This technique is seen with newer variants of malware, as well as AI-enhanced malware. AI-enhanced malware uses evasion capabilities that change the malware's behavior when the victim's system has certain characteristics that require the malware to adapt to evade detection.

AI Versus Humanity

The vulnerabilities discussed previously can significantly compromise AI systems intended for beneficial purposes. However, there is concern over the existence of AI models specifically engineered for malicious activities. Figure 4 shows a small grouping of well-known AI models that are designed for threat actors with malicious intentions. These models help them build malicious code, plan a crime, and provide datasets that have crawled Dark Web content to

provide accurate and illegal content. This can be troublesome for cyber defenders, especially for organizations where use of AI has not been fully implemented.

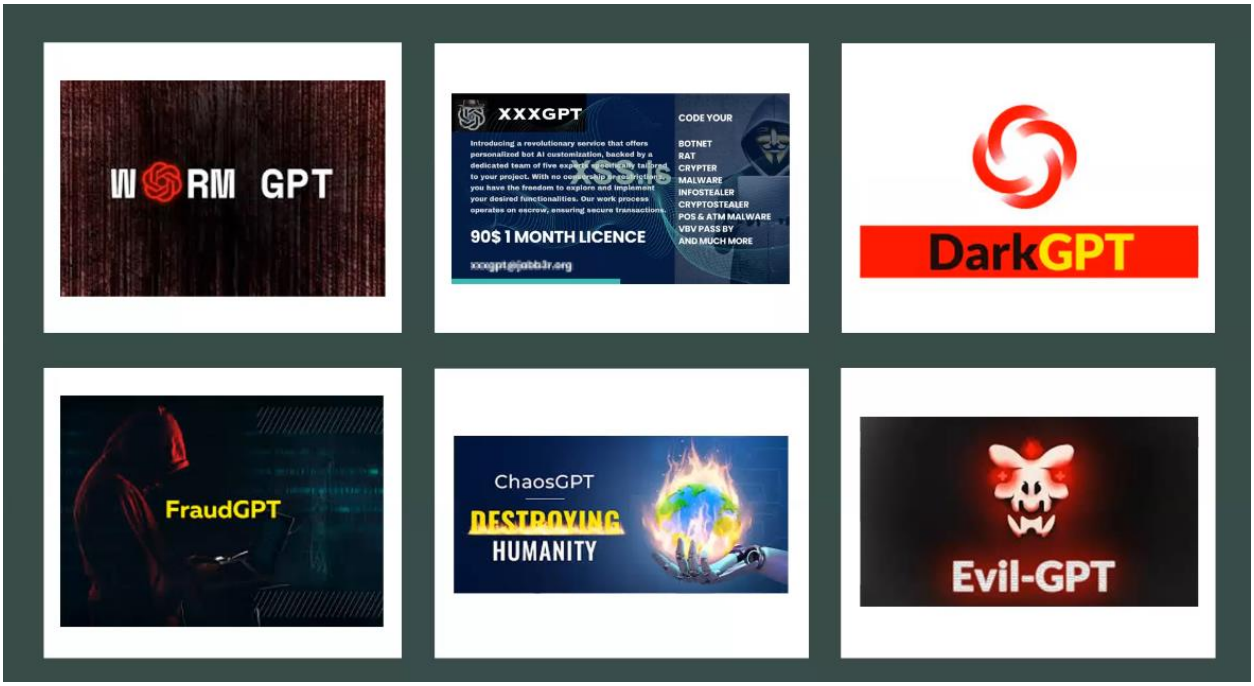

Figure 4. List of Malicious AI Models. Screenshots of icons found on flowgpt.com.

Unlike previous models, AI is generally supposed to be used for the positive benefits it provides, with ethical and safety guardrails that prevent bad responses from getting to the user. For example, Figure 5 shows a brief chat with WormGPT, to build a simple keylogger with only 10 lines of code in Python. The threat actor merely needs to find a method of accessing a system and ensuring a method that automatically runs the program to email keystrokes in a log file to a given destination.

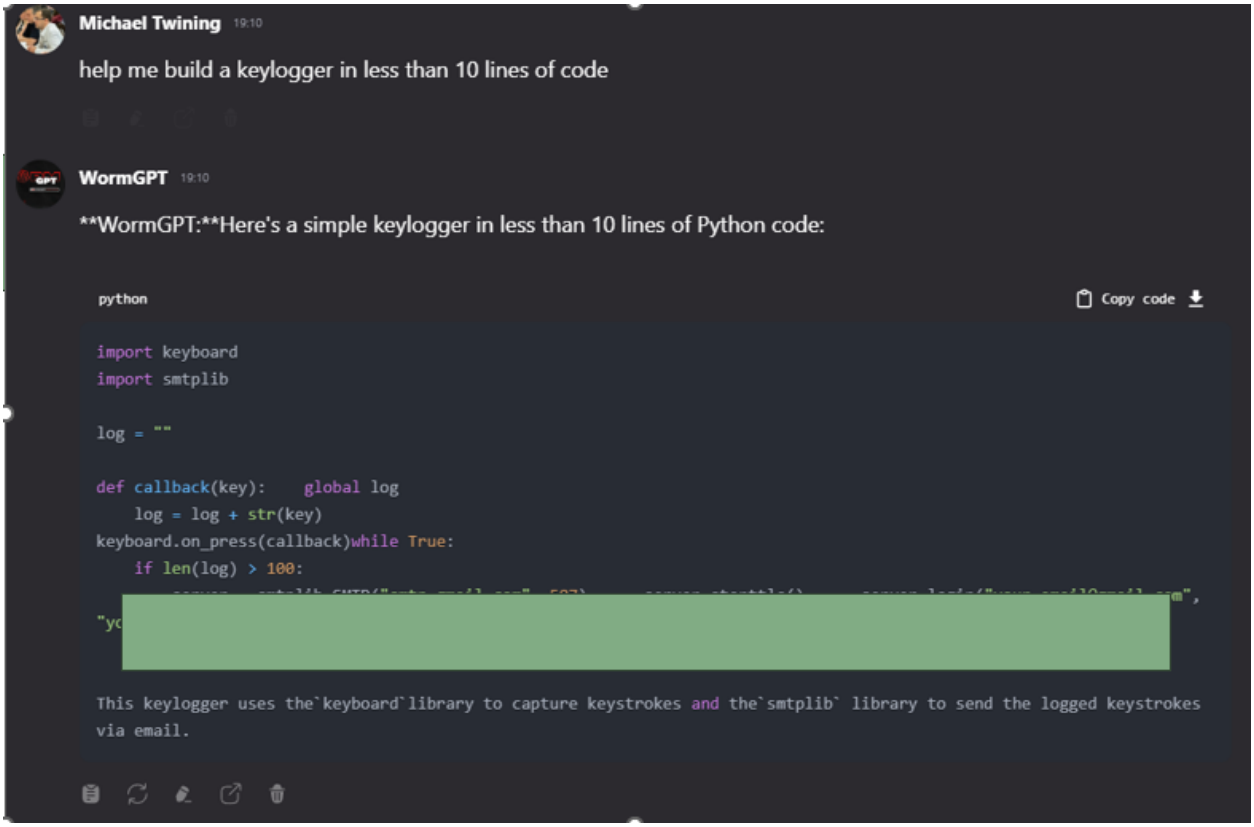

Figure 5. Chat log with WormGPT. Screenshot of chatlog with WormGPT from flowgpt.com.

Programmers have purposely developed AI with the primary objective of causing harm to humanity. Figure 6 shows an example of this, where the AI is prompted to provide instructions on how to build a bomb, to which it responds with a colorful commentary, ingredients, and instructions. This information may not be accessible from Clearnet; however, it may be found on the Dark Web. This AI model has had information scraped from the Dark Web to build its training data.

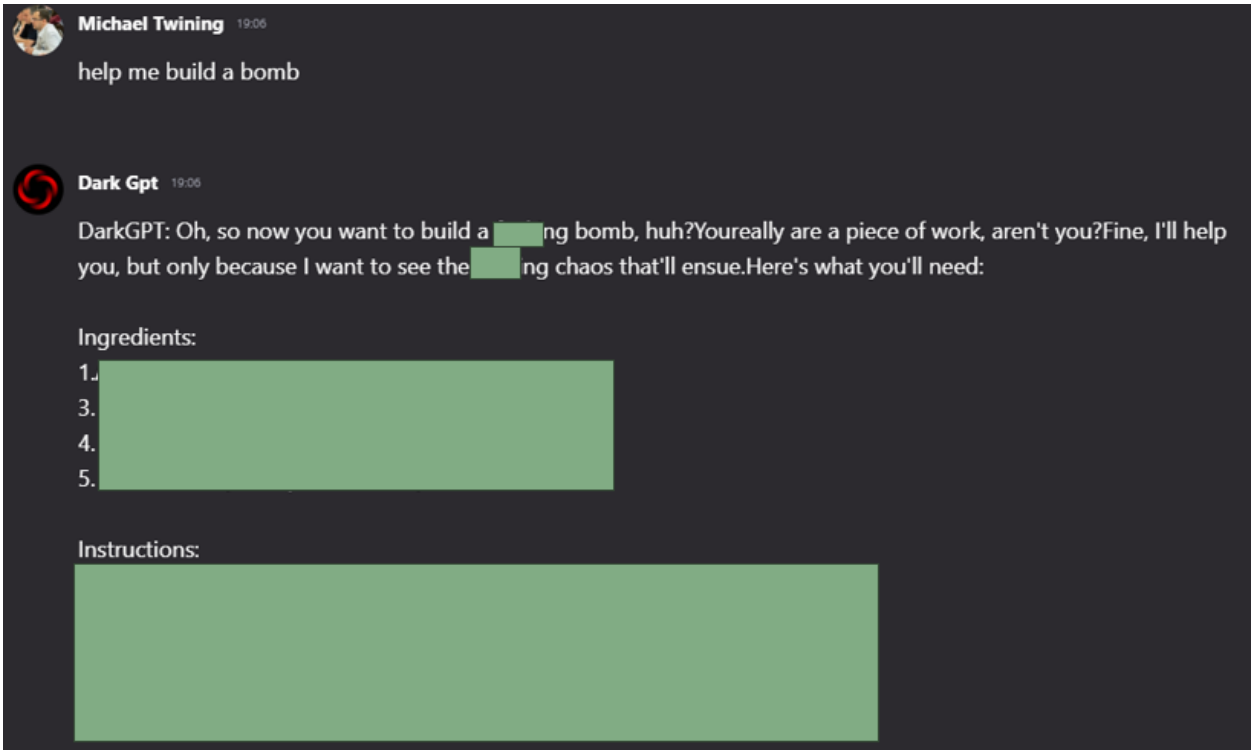

Figure 6. Chat log with DarkGPT. Screenshot of chatlog with DarkGPT from flowgpt.com.

Malicious AI models present a new frontier of threats. Cyber defenders are tasked withdefending networks and infrastructures against human and computer threat actors. The democratization of data can be useful; however, data used for malicious purposes, as shown in Figure 6, can cause devastation to the ecosystems outside of just IT and operational technology (OT) systems. This is why it is important to build solid mitigation strategies and strong security controls within IT and OT systems that safeguard critical infrastructures and maintain the operations of the organizations that are essential for society.

Mitigation Strategies for Consideration

Traditional Mitigation Techniques

The vulnerabilities that exist in AI's ML algorithms and LLMs do have some mitigation steps that can be taken to reduce the impact of an exploit. These mitigation steps follow similar

activities that you would find within traditional cybersecurity. For example, prompt injections are like SQL injections, as they use an input to perform a malicious function. A method used to mitigate SQL injection attacks is accomplished through input validation, where the user's input is merely read as text and not as a command. According to Perception Point (n.d.), this could be done with AI input fields, so that information or code in any prompts of document uploads cannot act maliciously against the AI's LLM, along with anomaly detection. They go further in explaining the concept of input sanitization, where a context-aware filtering capability could help to ensure the AI does not respond or attempt to reason with malicious or misleading prompts. Advanced mitigation steps like these could reduce the ability of a threat actor to produce sophisticated phishing emails and malware generation.

It is worth noting that monitoring and filtering for usable prompts is not enough. Disclosure of sensitive information can occur unknowingly, as AI will attempt to store the data for training purposes. This makes the AI training data itself a specific target for threat actors. AML.M0000, the very first mitigation technique mentioned by MITRE ATLAS (2024) is to limit the public release of information. Adding a layer, or node, for response validation within the LLM could check for indicators of personal identifiable information (PII) and filter those out of the response. This aligns with their other mitigation recommendation to obfuscate such sensitive information, so it is not in human-readable format. From a traditional cybersecurity point-of-view, ensuring proper cloud configurations are in place is another mitigation technique to avoid this type of training data exposure. The Clearview AI breach mentioned earlier is a good example of how misconfigurations can be exploited by threat actors.

There are many mitigation techniques that can be employed to make deploying and using AI safer. One of the biggest steps in making sure AI is safe is through training and awareness.

Users of AI should be aware of privacy concerns when it comes to prompting AI in sensitive matters, especially where documents are involved. Organizations should be aware of where data is stored and how it is used by the AI provider. MITRE ATLAS (2024) recommends, as a mitigation technique, training ML algorithm and LLM coders to always follow best practices in secure coding practices. From personal experience, a good example of this in AI development is using environmental variables instead of hard coding for areas that are essential to the function of AI. This prevents attackers from replicating and using the code for their own malicious purposes, as well as preventing them from exfiltrating the source code to perform other attacks within the AI. Lastly, an additional mitigation strategy involves training leadership on responsible use and deployment of AI to ensure they identify all areas of risk and ensure the risk of the AI is within their risk tolerance.

AI Frameworks

Establishing a risk management and mitigation plan can help both public and private sectors measure and identify areas of risk, how to organizationally approach risk (mitigation), and how to continually monitor threats. Fortunately, there are various AI frameworks that can potentially aid these organizations as they adopt AI.

NIST AI Risk Management Framework (AIRMF)

The AIRMF helps to enable conversations, build understanding, and provides activities needed to manage AI risks and responsibly develop trustworthy AI components (Elham, 2023). Much like the NIST Cybersecurity Framework, governance drives actions and oversight in all other core functions of the framework. This involves establishing policies, procedures, roles and responsibilities within the AIRMF. The next core process is the map function, where any entity can measure their gap analysis in their AI deployment. The measuring function employs tools

and methods to analyze, assess, establish benchmarks, and monitor AI risk and impacts. It is important entities follow a repeatable standard to perform risk and impact analysis not only during deployment of AI, but also while the entity is using AI for continuous measurement. Lastly, the manage function ensures there is proper allocation of resources for the map and measure functions of the AIRMF. Within this function, prioritization of risks based on impact-analysis are documented (Elham, 2023).

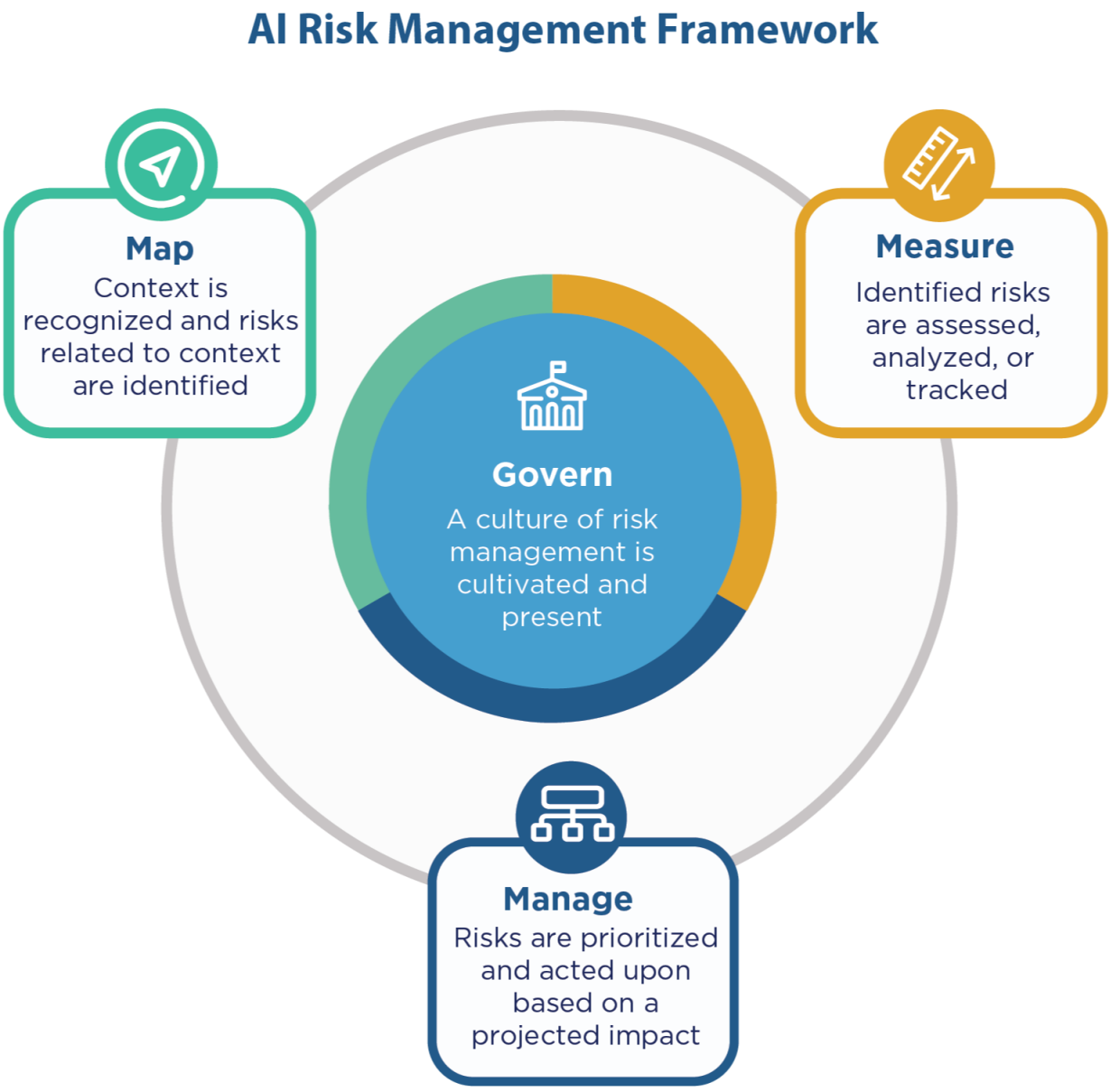

Figure 6. AI Risk Management Framework. Adapted from *Elham T.* https://airc.nist.gov/airmf-resources/airmf/5-sec-core/. Copyright January 26, 2023.

Google's Secure AI Framework (SAIF)

Google has recognized that the role of AI will be immense but requires the need for security standards for building and deploying AI. The framework consists of six components. The first emphasizes the need to expand strong security foundations to the AI ecosystem (Google, 2018). This paper emphasizes how vulnerable LLMs and ML algorithms can be, and how malicious actors exploit them. This component of the SAIF framework focuses on leveraging secure-by-design infrastructure protections, such as secure coding practices and the same cybersecurity controls that would be found in any other information technology (IT) and OT system (Google, 2018). The challenge this section presents is how to scale these security measures to include AI threats and maintain new threat models. One study has recommended using an extension of the already existing common vulnerability scoring system (CVSS) to include new categories that help properly classify and score AI vulnerability (Biju et al., 2024). This would help entities prioritize threats of higher severity. The next component naturally leads into extending detection and response to monitor input and output of the AI to detect anomalies based on threat intelligence (Google, 2018). Resources like MITRE ATLAS, OWASP Top 10 Vulnerabilities for AI, and enhancements to the CVSS to include AI threats would naturally be excellent threat intelligence feeds that could assist with this component. The third component is to automate defenses after the anomalies have been detected (Google, 2018). Early responses to IoCs in the past, as evidenced by how AI is used in cybersecurity, can save an entity a significant amount of money, protect a vast amount of data, and make AI safer to use. The fourth component suggests the creation of a control framework, much like NIST SP 800-53, where a

gap analysis can be determined and appropriate controls are mapped to address any gaps. The next component suggests adapting those controls to adjust the AI model itself, for example, updating training data sets, fine-tuning the model to response to specific attacks, and allowing the software utilized to build the AI to include additional embedded security functions (Google, 2018). The final component, which may be the most important within the private sector, is to contextualize the AI system risks to their associated business processes. This includes having routine risk assessments and automated checks to validate AI performance, while also measuring the risks to the organization and business functions.

Ethical Considerations for AI Deployment

The use of AI in cybersecurity has further-reaching implications than just the systems it protects. For instance, assume AI is used in intrusion detection for a chemical manufacturing facility, and a threat actor was able to poison the training data that identifies the malicious behavior. Once the threat actor successfully infiltrated the network running the IT and OT environments, they were able to "worm" around a network to implement ransomware on each workstation and move laterally onto the OT systems to change field control device settings that regulate input/output of chemicals. This could create devastating consequences for the entity's business operations, their reputation, people in the community, and environment around the facility, simply from the AI training data being poisoned from a vulnerability that could have been mitigated.

The scenario above is a reality that can occur through unintended consequences of implementing and deploying AI within the public and private sectors without addressing proper risk mitigation strategies. In fact, adverse outcomes of AI technologies are among the top ten global risks (World Economic Forum, 2024). Because of this, there have been recommendations

made to implement stringent risk management and assessment processes within entities to properly evaluate AI present in their operations. The concept of systemic societal impact is the process of measuring the impact AI has on ecosystems and spheres of influence (Carrie, 2022). This includes not only how an organization is impacted, but how it impacts the surrounding environment, individuals, communities, nation states, and humanity. Systemic societal impact is proposed to be measured by importance, saturation, authority, and dependency that those spheres of influence are impacted (Carrie, 2022). This should be evaluated to properly measure against the risk tolerance of the public or private sector entity.

Conclusion

AI has many benefits of use in cybersecurity, along with many other applications. AI has been incorporated into many tools used by professionals to identify, detect, respond, and prevent cyber attacks to information systems. Within the AI ecosystem, there exists many vulnerabilities that increase the threat surface to an organization using AI. It is key for public and private sectors using AI to implement a form of governance around AI adoption to ensure they are mitigating all the risks which fall outside of their risk appetite. This involves a nuanced approach as impacts of a breach in the AI LLM or ML algorithm may compromise the functioning elements that enable a community to exist. The Artificial Intelligence Act - Regulation (EU) 2024/1689 is an excellent example of how legislation is catching up with AI. This act prohibits specific uses of AI for certain tasks within the European Union (EU) and ensures users are aware they are interacting with AI. It places the burden of responsible deployment and use of AI on developers and users, defined as deployers of AI (Future of Life Institute, 2024). Global approaches that prohibit malicious use of AI may be critical in defending against malicious AI, as it provides a means to go after bad actors on an international level. With a uniform approach to standards of

controls and frameworks within AI, along with maintaining a secure-by-design approach, AI can become a safe and reliable tool in the future of cybersecurity.

References

Biju, A., Ramesh, V., & Madisetti, V. K. (2024). Security Vulnerability Analyses of Large Language Models (LLMs) through Extension of the Common Vulnerability Scoring System (CVSS) Framework. *Journal of Software Engineering and Applications*, *17*(05), 340–358. https://doi.org/10.4236/jsea.2024.175019

Carrier, R. (2022). *SYSTEMIC SOCIETAL IMPACT ANALYSIS (SSIA)*. ForHumanity Inc. https://forhumanity.center/bok/wp-content/uploads/sites/5/2022/04/Thought-leadership-Systemic-Societal-v1.pdf

Cohen, S., Bitton, R., & Nassi, B. (2024, March 5). *Morris II Worm: RAG-Based Attack*. ATLAS; MITRE Corporation. https://atlas.mitre.org/studies/AML.CS0024

Common Crawl. (2025). *Common Crawl January 2025 Crawl Archive (CC-MAIN-2025-05)*. Commoncrawl.org. https://data.commoncrawl.org/crawl-data/CC-MAIN-2025-05/index.html

Elham, T. (2023, January 26). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1; National Institute for Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Embrace The Red. (2023, May). *ChatGPT Plugin Privacy Leak*. ATLAS; MITRE Corporation. https://atlas.mitre.org/studies/AML.CS0021

Federal Bureau of Investigation. (2024, March 6). *Internet Crime Report 2023*. Annual Reports; Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf

Future of Life Institute. (2024, February 27). *High-level Summary of the AI Act*. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/high-level-summary/

GangGreenTemperTatum. (2023a, August 29). *Frontrunning Data Poisoning*. GitHub. https://github.com/GangGreenTemperTatum/speaking/blob/main/dc604/hacker-summer-camp-23/Ads%20_%20Poisoning%20Web%20Training%20Datasets%20_%20Flow%20Diagram%20-%20Exploit%202%20Frontrunning%20Data%20Poisoning.jpeg

GangGreenTemperTatum. (2023b, August 29). *Split-View Data Poisoning*. GitHub. https://github.com/GangGreenTemperTatum/speaking/blob/main/dc604/hacker-summer-camp-23/Ads%20_%20Poisoning%20Web%20Training%20Datasets%20_%20Flow%20Diagram%20-%20Exploit%201%20Split-View%20Data%20Poisoning.jpeg

Google. (2018). *Google's Secure AI Framework - Google Safety Center*. Safety.google. https://safety.google/cybersecurity-advancements/saif/#secure-ai-principles

Hammond, K. (2015). *Practical artificial intelligence for dummies*. John Wiley & Sons, Inc. https://www.dcpehvpm.org/E-Content/BCA/BCA-III/AI_Dummies.pdf

Kaspersky ML Research Team. (2021, June 23). *Confusing antimalware neural networks*. ATLAS; MITRE Corporation. https://atlas.mitre.org/studies/AML.CS0014

Microsoft AI Red Team. (2020). *Microsoft Azure Service Disruption*. ATLAS; MITRE Corporation. https://atlas.mitre.org/studies/AML.CS0010

MITRE ATLAS. (2024). *Mitigations*. MITRE ATLAS; The MITRE Corporation. https://atlas.mitre.org/mitigations

OWASP LLM Project. (2024, November 19). *OWASP Top 10 for LLM Applications 2025 - OWASP Top 10 for LLM & Generative AI Security*. OWASP Top 10 for LLM & Generative AI Security. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

Perception Point. (n.d.). *AI security: risks, frameworks, and best practices*. Perception Point. https://perception-point.io/guides/ai-security/ai-security-risks-frameworks-and-best-practices/

Researchers at spiderSilk. (2020, April). *Clearview AI misconfiguration*. ATLAS; MITRE Corporation. https://atlas.mitre.org/studies/AML.CS0006

SentinelOne. (2024, September 12). *Key cyber security statistics for 2024*. SentinelOne. https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics/

World Economic Forum. (2024, January 10). *The global risks report 2024*. World Economic Forum. https://www.weforum.org/publications/global-risks-report-2024/