Emerging Threats in Artificial Intelligence[1]

Dominick Zangaro, Louis Fritz, Curtis Thomas, Kaiden Brittain
Cyber Risk Division

Introduction

Whether your organization is adopting Artificial Intelligence (AI) or not, AI will fundamentally change the cyber security landscape to how cyber criminals conduct their business. While not a comprehensive list, we wanted to outline multiple ways that AI can be leveraged to inflict financial harm, cause computer outages, and destroy your organizations' reputation. Below, we have outlined those risks, with Large Language Models (LLMs) and Generative AI (GenAI) at the forefront, followed by the expanding capabilities of Malware (ransomware) and Distributed Denial of Service (DDoS) attacks.

What Is AI?

To begin, we wanted to outline what AI is, how it works, and the value that AI provides to benevolent users of the technology. AI is the simulation of human intelligence wherein machines are taught to think and learn as humans would. There are four tiers of AI that exist today with categories that depend on a model's capabilities and how close it is to replicating human intelligence perfectly. The lowest tier is called Reactive Machines as this form of AI is designed to react to current situations or inputs with no memories to aid in forming future decisions. Limited Memory builds off Reactive Machines by incorporating memories into the decision-making process, like the work being done today with autonomous vehicles. Theory of Mind AI is still in development and is actively being designed to allow computing technology to understand human emotions, beliefs, and thoughts to deepen its knowledge and connect more effectively with humans. Finally, Self-Aware AI is the most

[1] Any views or opinions expressed in this paper are the authors' and shall not be construed as legal or professional technical advice; and do not necessarily reflect any corporate position, opinion or view of Great American Insurance Company, or its affiliates, or a corporate endorsement, position or preference with respect to any contractual terms and provisions or any related issues. If you have any questions or issues of a specific nature, you should consult appropriate legal or regulatory counsel to review the specific circumstances involved. Cyber loss control consultation services are provided by Great American Insurance Company and its affiliates to assist management of insured firms in fulfilling their responsibilities for the control of potential loss producing situations involving their information technology and/or operations. The information provided is intended to provide guidance and is not intended as a legal interpretation of any federal, state or local laws, rules or regulations applicable to your business. The cyber loss control information provided is intended only to assist policyholders in the management of potential loss producing conditions involving their information technology and/or operations based on best practices around cybersecurity controls. In providing such information, Great American does not warrant that all potential hazards or conditions have been evaluated or can be controlled. It is not intended as an offer to write insurance for such conditions or exposures. The liability of Great American Insurance Company and its affiliated insurers is limited to the terms, limits and conditions of the insurance policies underwritten by any of them.
© 2025 Great American Insurance Company, 301 E. Fourth St., Cincinnati, OH 45202. All rights reserved.

advanced tier and is currently theoretical. This type of AI will possess both self-awareness but also human consciousness.[2]

AI leverages a suite of technologies, with each bringing its own value, and depending on the AI Model, blends each in a unique way. These core technologies may include Machine Learning (ML), which is designed to train algorithms on data sets to learn and make decisions, Deep Learning, which expands on ML by introducing neural networks within computing technologies (mimicking the physical structure of the human brain), Natural Language Processing (NLP), which allows technology to understand, interpret, and respond to human language, and Computer Vision, which is an input of visual data that is analyzed and acted upon by AI.[3]

To a benevolent user, these tools present a unique opportunity to automate manual processes, develop new products, and analyze troves of data efficiently. The power that AI presents to businesses is immense, as value can be unlocked in new and exciting ways, yet unimagined. For example, AI can be used to analyze sales data across industries and geographic regions in near real time. Also, IT professionals may use it to identify bottlenecks in a network, exhaust computing capabilities on individual devices across an environment, and security teams may leverage AI to analyze security alerts whether physical or cyber in nature. These efficiencies can become a competitive advantage for organizations of all shapes and sizes, and further AI adoption will continue throughout the decade.

AI poisoning

Now that we have familiarized ourselves with AI, how it works, and why its potential value to organizations is growing exponentially, understanding how it can be used to cause harm is diligent cyber risk management. The first risk, AI Poisoning, is related to Generative (GenAI) AI and Large Language Models (LLMs) and how data being input can alter future outputs. Specifically, but not in every case, an individual with malicious intent might compromise an AI resource by injecting false data sets that modify existing data, or that entirely removes datasets from being analyzed.

Whether you are implementing AI into your organization's environment or developing your own model, data injection is just one exposure an attacker may be able to leverage for financial gain. Backdoor attacks are just as much a problem, as a sophisticated user, rather than inject false data, may craft a prompt that causes an AI model to remove security controls, identify gaps in their security, and even create vulnerabilities that can be used to gain root level access. Lastly, an emerging area of concern is the stealth attack, which functions very similarly to both AI poisoning and Backdoor Attacks. Rather than trying to quickly alter an AI model, stealth attacks take a patient approach. Over time, an attacker will incorporate small errors into prompts, or inputs, which are hard to detect and only slightly change the way the model functions. In doing so, security analysts

[2] Coursera, *What Is Artificial Intelligence? Definition, Uses, and Types* (2024). https://www.coursera.org/articles/what-is-artificial-intelligence (last accessed November 14, 2024).
[3] *Id.*

will be faced with a more challenging task when identifying these errors and taking corrective action.[4] This can lead to false outputs, and tainted data sets that undermine the validity of your model.[5]

While these risks are still in their infancy and researchers are working diligently to understand each nuance, security experts have begun to identify a few initial cyber security controls that can mitigate the risks outlined above. One of those, Data Validation, is the process of continuously validating the classification labels and quality of data to ensure that it is accurate and representative of its intended use. In part, this creates security alerts when anomalous patterns emerge, that should not be ignored. In doing so, you can leverage another defense, grounded in the training of your AI model. Adversarial training is the idea that your model would be exposed to malicious inputs so that it can identify and mitigate the risk in real time. This improves the overall resiliency of the model and reinforces the overall accuracy and validity of outputs.[6] Finally, new business models are being developed that focus their products on securing AI and AI models. Recently, MITRE, responsible for the MITRE ATT&CK and D3FEND, and other organizations created an AI management framework called the Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) that encompasses 14 tactics used by attackers and how to defend against these attacks. Additionally, groups like Microsoft have launched open-source platforms (Microsoft Counterfeit) that conduct security assessments against AI models.[7]

Model Inversion Attacks

AI Model Inversion (MI) attacks are a significant threat in machine learning, particularly involving identity theft and model theft. These attacks allow adversaries to infer sensitive information from a trained model. Understanding these risks and implementing effective mitigation strategies is crucial, especially for industries like insurance that rely heavily on data-driven decision-making.

A Model Inversion attack involves an adversary gaining access to a machine learning model to reconstruct input data from the model's outputs. Inputs are the data points or features that are used to train the model. Outputs are the results produced by the model after processing the inputs. By exploiting the model's learned parameters and the relationships between inputs and outputs, attackers can query the model and use the responses to infer sensitive information about the training data.[8]

[4] Richard Jenkins, *Defending the future: a guide to fortifying AI against data poisoning attacks* (2023). https://www.glasswall.com/blog/defending-the-future-a-guide-to-fortifying-ai-against-data-poisoning-attacks (last accessed November 14, 2024).

[5] Wiz Experts, *What is a Data Poisoning Attack?* (2024). https://www.wiz.io/academy/data-poisoning (last accessed November 14, 2024).

[6] Maria Korolov, Adversarial machine learning explained: How attackers disrupt AI and ML systems, CSO Online (2022). https://www.csoonline.com/article/573031/adversarial-machine-learning-explained-how-attackers-disrupt-ai-and-ml-systems.html (last accessed November 14, 2024).

[7] *Id*.

[8] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li and Yi, Yang, *Random Erasing Data Augmentation*, Proceedings of the AAAI Conference on Artificial Intelligence (2020).

The risks associated with MI attacks are substantial. In terms of identity theft, MI attacks can reconstruct confidential data, such as personal images or financial details. For instance, in facial recognition systems, by using AI, attackers can recreate images of individuals used in the training data. This reconstructed data can be used for malicious purposes, including identity theft, fraud, and unauthorized access to personal accounts.[9] Regarding model theft, attackers can reverse-engineer the model's parameters and architecture, effectively stealing the intellectual property. This allows them to replicate and misuse the model without authorization, putting the original developer at a competitive disadvantage.

To mitigate these risks, several strategies can be employed. Differential privacy involves adding noise to the training data or model outputs to obscure sensitive information, making it harder for attackers to reconstruct accurate data. Adversarial training involves training models with adversarial examples to make them more resistant to MI attacks, enhancing their overall security. Access control is crucial, limiting access to the model and its outputs to trusted users only and regularly monitoring and auditing model usage to detect any unusual activity. Random erasing techniques during training can degrade the quality of data reconstruction by MI attacks without significantly affecting model performance. Mutual information regularization helps minimize the amount of confidential information encoded in the model, balancing the trade-off between model utility and privacy.[10]

In conclusion, Model Inversion attacks pose significant risks to the privacy and integrity of machine learning models.

Mutating Malware and the Ransomware Problem

Before digging into mutating malware and its capabilities, it is important to understand what Malware is and the many forms that it takes. Malware is a 'catch all' term for any software that is intentionally malicious in nature; that causes harm to a computer, network, or environment. Many common types of malware include, but are not limited to, computer viruses, spyware, ransomware, and trojan horses. Malware works to infect these same systems through phishing emails, system vulnerabilities and infected files to name a few.

Historically, we have defended against these threats by taking a defense in depth approach to cyber security. This thinking is like that of a castle with many walls defending against repeated attacks. Creating strong passwords, using anti-virus and firewalls, implementing comprehensive endpoint detection and response, and staying current with all patches and updates can mitigate the risk of compromise by malware. The introduction of Artificial Intelligence, greatly changes this dynamic as malicious actors can now incorporate this technology to bypass these controls and are better suited to exploit the human element of cyber security by impersonating

[9] Michael Tschannen, Josip Djolnga, Paul Rubenstein, Zylvain Gelly and Mario Lucic, *On Mutual Information Maximization for Representation Learning*, International Conference on Learning Representations (2020).

[10] Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy,* Foundations and Trends in Theoretical Computer Science (August 11, 2024).

trusted stakeholders, creating urgency, and playing on fear in new and exacerbating ways. This section will detail these emerging risks and their potential impacts.

Mutating Malware is defined as malicious software that is designed to evade detection by changing its appearance, code, or set of behaviors in real time. The two driving types of mutating malware are Polymorphic and Metamorphic Malware. Polymorphic Malware alters its code on each infection by changing its decryption routine while maintaining its core functionality.[11] This creates a security gap as traditional anti-virus, which relies on signature-based detection, can no longer identify and block malware execution. Metamorphic Malware takes it a step further by altering the core functionality of its code, in addition to the decryption routine, which makes it harder for identification by behavior-based detection from Endpoint Detection and Response (EDR), Security Information and Event Management (SIEM), and Security Operations Center (SOC) teams.[12]

These nuances create significant risk for organizations, as traditional approaches to cyber security are no longer as effective in mitigating risk as they once were. These new tools allow a ransomware group to gain initial access with more ease, whether through AI-powered Social Engineering, or by exploiting weaknesses in current detection techniques. Once in, AI-powered malware, specifically ransomware, can alter its appearance to bypass additional layers of security while simultaneously scanning your environment to identify and target business critical systems, endpoints, and sensitive / confidential databases to inflict as much damage as possible. Today, ransomware groups need to do all of this on their own and may fail to bypass security and/or execute their payload on the most impactful systems. This means that there is time for humans to intercept malicious activity, develop plans to prop up less critical assets to continue operations, and allocate more resources to monitoring more critical assets. These steps have reduced the overall frequency and severity of ransomware, since 2022 (per the Verizon Data Breach Investigations Report).[13] While there has been relative stability with ransomware proliferation, these metaphoric "calm waters" will soon be choppy once again as these tools will unlock the capability for more catastrophic loss.

So, what can you do today to defend against the threats of tomorrow? Begin by reviewing your current detection and response tools to determine if they are currently tracking behaviors with AI-powered Threat Intelligence. This can be done in part with the use of AI itself through machine learning.[14] Also, the next generation antivirus leverages signature-less detection that can block malicious polymorphic and metamorphic malware.[15]

[11] Malwarebytes, *Polymorphic virus* (2024). https://www.malwarebytes.com/polymorphic-virus (last accessed November 14, 2024).
[12] Casey Crane, *Polymorphic Malware*, The SSL Store (2024). https://www.thesslstore.com/blog/polymorphic-malware-and-metamorphic-malware-what-you-need-to-know/ (last accessed November 14, 2024).
[13] Verizon, *Data Breach Investigations Report* (2022). https://www.verizon.com/business/en-gb/resources/2022-data-breach-investigations-report-dbir.pdf (last accessed November 14, 2024).
[14] Coursera, *supra*, note 1.
[15] Jenkins, *supra*, note 3.

AI Driven DDoS

A Distributed Denial of Service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service, or network by overwhelming it with a flood of internet traffic. This is achieved by using multiple compromised computer systems as sources of attack traffic, often forming a botnet. The sheer volume of incoming messages, connection requests, or malformed packets to the target system forces it to slow down or crash, rendering it inaccessible to legitimate users. DDoS attacks can cause significant downtime and economic loss for businesses and are a common tool for cybercriminals to disrupt services or extort funds.

All organizations are targets for DDoS. These attacks are often used to extort businesses dependent on technology, or depending on the year, and are used by Hacktivists against media or political groups. Telecom companies are prime targets for DDoS attacks because they play a crucial role in global connectivity, making any disruption highly impactful. Telecommunications was the second most targeted industry, but it experienced the largest surge in attacks, with a fivefold increase in 2023 compared to the prior year.[16] Telecoms are juicy targets as a successful attack can cause widespread outages, affecting millions of users and businesses. It is no surprise that these are highlighted from a ransomware scenario perspective as they hold the most "weight" from a support perspective from the top down.

Historically, defending against DDoS attacks has involved a combination of preventive measures like regular software updates and good cybersecurity practices, packet filtering to block malicious traffic, and rate limiting to prevent any single source from overwhelming servers. DDoS attacks, have been mitigated through various hardware solutions are, including Web Application Firewalls (WAF) and Intrusion Prevention Systems (IPS) that detect and block malicious traffic, and load balancers that distribute traffic across multiple servers to prevent overload. DDoS from an attacker's perspective is a race against time and are executed in the attempt to prohibit network function through overwhelming force at the expense of in-depth network understanding and the inconspicuousness of attacks that take more time to roll out.

AI is being used to dynamically learn a target's network structure and traffic patterns to determine the optimal time, place, and method in which the attacker should strike. Information can be gathered by AI scrapers that work through websites and through AI engine poisoning so that connection is made and sent from an open AI address is mistaken for a standard network transaction or interaction to reduce the chances of detection.[17] Additionally, rogue chatbots or plugins that leverage chatbots can be used to monitor interactions from users of target domains and can gather information on their network.

[16] David Warburton and Malcom Heath, *2024 DDoS Attack Trends*, F5 Labs (July 16, 2024). https://www.f5.com/labs/articles/threatintelligence/2024ddosattacktrends#:~:text=That%20same%20organization%20also%20suffered,month%20reaching%20less%20than%2010Gbps (last accessed November 14, 2024).

[17] David Aviv and Mark Hopt, *Is AI a factor in DDoS Attacks*: CISO Corner Podcast, Episode 26 (2024). https://www.databank.com/resources/videos/is-artificial-intelligence-a-factor-in-ddos-attacks/ (last accessed November 14, 2024).

The largest factor is understanding the structure of your website. If I, as an attacker, know the structure, I can build a targeted DDoS attack to hit you in a way that is tailored to your weak points. Additionally, now with the use of AI, attackers can shift or evolve attacks based on defense strategies in use. If an attacker started with a SYN flood attack (sending overwhelming packets) that at a certain point is not making any ground or progress on their end, they can change gears and launch an attack on Layer 7 (application Layer).

At this point, it is not enough to just utilize hardware to mitigate DDoS attacks. It is not a question of whether you will be attacked, but when. With organizations of varying scale, not everyone is a giant like Google and can defend based on their own resources.[18] Today, there are also tools and services that utilize monitoring and have humans in the loop. It is important to remember that AI is just machine learning and the utilization of live statistical analysis, and these same functions exist on the side of the good guys also but take time to adapt and implement. Organizations are testing functions of AI usage within defense software and toolsets to help combat the evolving landscape of DDoS.

[18] Anna Claiborne, *AI and the rise of DDoS Attacks*: Infosec Podcast (2024). https://www.infosecinstitute.com/podcast/ai-and-the-rise-of-ddos-attacks--guest-anna-claiborne/ (last accessed November 14, 2024).