Strengthening AI Critical Infrastructure Security with the MIT AI Risk Repository and MITRE ATLAS Frameworks

Jami Carroll
Harvard University, Somerset, MA, USA

jcarroll@prisidian.com

Abstract: Artificial Intelligence (AI) plays a pivotal role in critical infrastructure sectors such as energy, finance, healthcare, defense, and transportation. These sectors benefit from AI's advanced capabilities, including predictive analytics, automation, and enhanced decision-making. However, AI integration also introduces significant security risks, such as adversarial attacks, data poisoning, and vulnerabilities within supply chains, potentially leading to system compromise and operational failures. Addressing these challenges requires a structured and proactive risk assessment approach. This study proposes a comprehensive AI security framework leveraging the MIT AI Risk Repository, which consolidates 43 frameworks, 2 taxonomies, and 777 identified risks, and MITRE ATLAS, which documents over 1500 attack vectors against AI systems. A systematic review of AI security research from 2020-2024 was conducted to assess common attack vectors, including deepfake technology, AI system poisoning, and supply chain threats. By mapping AI vulnerabilities to adversarial tactics, this research provides a structured methodology for identifying and mitigating risks. The findings contribute to establishing robust cybersecurity practices, enhancing AI resilience, and guiding policy development for critical infrastructure protection. This study highlights the importance of adopting AI-specific security frameworks to mitigate emerging threats and safeguard AI-driven systems across industries.

Keywords: AI security, Critical infrastructure, Vulnerability assessment, Cybersecurity practices, Risk mitigation, MIT AI risk repository, MITRE ATLAS

1. Introduction

The integration of Artificial Intelligence (AI) into critical infrastructure sectors—such as energy, finance, healthcare, defense, and transportation—has delivered transformative advancements, significantly enhancing operational efficiency and service capabilities. However, the widespread adoption of AI technologies introduces complex and evolving cybersecurity risks, highlighting vulnerabilities such as deepfake technologies, AI system poisoning, and supply chain attacks. Addressing these sophisticated threats requires a comprehensive, structured approach to risk assessment and mitigation. This paper explores critical vulnerabilities associated with AI deployment and emphasizes the strategic benefit of employing established frameworks such as the MIT AI Risk Repository and MITRE ATLAS. By adopting these frameworks, organizations can systematically enhance their AI security posture, ensuring resilience against emerging and sophisticated cyber threats.

2. Research Criticality

The rapid integration of Artificial Intelligence (AI) into various sectors, especially critical infrastructures, significantly amplifies the urgency to strengthen AI platforms against increasingly sophisticated and rapidly evolving security threats. Three key areas — deepfake technology, AI system poisoning, and supply chain security—represent particularly pressing vulnerabilities requiring structured assessment, timely identification, and effective mitigation strategies.

Deepfake Technology

Deepfake technology, powered by generative AI, presents significant security challenges. Initially popularized in film production for creating digital impersonations, its implications extend far beyond entertainment. These digital creations are now sophisticated enough to bypass systems of identification, authentication, and authorization, facilitating identity theft, fraud, and extensive misinformation campaigns posing severe security risks. By leveraging advanced machine learning techniques, particularly Generative Adversarial Networks (GANs), deepfake technology enables the creation of convincing synthetic media capable of bypassing traditional authentication systems. The accelerated advancement of GANs and related generative models continues to enhance the realism and sophistication of deepfakes, drastically complicating their detection and control. Current detection tools rely heavily on identifying inconsistencies such as unnatural lighting, blurred edges, or unusual angles; however, rapid improvements in GAN capabilities frequently outpace detection methods (Masood et al., 2023; Murphy, 2023). As deepfake instances have surged from approximately 15,000 in 2019 to millions in recent years, organizations urgently require sophisticated, adaptive security mechanisms to manage and counteract the potential damage posed by this technology

AI System Poisoning

AI system poisoning, another critical security concern, involves the subtle manipulation of datasets utilized to train AI models, resulting in compromised AI performance and reliability. Attackers may strategically introduce corrupted or misleading data points into AI systems, significantly impacting outcomes even with minimal alterations costing as little as little as $60 USD and affecting 0.01% of a dataset to severely hamper the AI system. This form of attack poses severe risks, particularly in high-stakes environments like financial institutions, healthcare providers, and defense sectors, where reliability and accuracy are paramount. Highlighted by recent research demonstrations, including work presented by Nicholas Carlini at Google Brain, these attacks reveal the alarming ease with which malicious actors can subtly but effectively compromise AI models (Carlini et al., 2024; Tong, 2023). The complexity and subtlety of these attacks underscore the critical need for robust defensive measures and ongoing vigilance in monitoring AI system integrity.

Supply Chain Security

Supply chain security emerges as an equally vital area of concern, given the reliance of AI systems on an extensive network of interconnected hardware and software suppliers. High-profile cyber incidents, including the SolarWinds, Colonial Pipeline, and JBS USA attacks during the early-2020s, underscore the profound and far-reaching implications of supply chain vulnerabilities (Biden, 2021; NSA, 2022a; NSA, 2022b; NSA, 2022c). While software is a key attack vector, Gnad et al. (2024) points out that vulnerabilities in both hardware and software essential for AI operations could lead to unauthorized data access or system manipulations, including side-channel attacks. These incidents vividly demonstrate how breaches within the supply chain can disrupt critical services and operations, lead to substantial economic damage, and facilitate unauthorized access or malicious manipulation of sensitive AI-driven systems. Such threats highlight the expansive attack surface created by intricate supply chains and the necessity for comprehensive risk management strategies to protect AI-enabled operations from potential exploitation.

Benefits of the MIT AI Risk Repository and the MITRE ATLAS Frameworks

To address these complex security challenges, organizations are encouraged to leverage established resources such as the MIT AI Risk Repository and the MITRE ATLAS frameworks. The MIT AI Risk Repository offers a comprehensive consolidation of 43 frameworks, 2 taxonomies, and over 777 identified AI-specific risks, providing a structured foundation for identifying, evaluating, and mitigating AI-related vulnerabilities. Complementing this, the MITRE ATLAS catalogs over 1500 documented attack vectors, offering detailed insights into potential adversarial actions targeting AI systems (MIT, 2025; MITRE, 2024a; MITRE, 2024b; MITRE, 2024c). By utilizing these frameworks, stakeholders gain a structured methodology for systematically assessing risks, recognizing vulnerabilities, evaluating potential impacts, and implementing effective mitigation strategies. Ultimately, integrating these robust frameworks into organizational security protocols significantly enhances the resilience of AI platforms, providing proactive measures to counteract evolving cyber threats and securing critical AI-dependent infrastructures.

3. AI Applications Across U.S. Critical Infrastructure: Enhancing Security and Operational Resilience

Artificial Intelligence (AI) has been strategically integrated across various critical infrastructure sectors in the United States, playing an essential role in augmenting both the efficiency and security of operations. Table 1 outlines the diverse sectors considered critical infrastructure, highlighting areas such as Energy, Healthcare, Transportation Systems, Defense Industrial Base, and Financial Services, among others. Each sector uniquely benefits from the implementation of AI, as detailed in Table 2, showcasing AI's substantial contribution to optimizing operational efficiencies, improving predictive capabilities, and bolstering cybersecurity measures.

In the energy sector, AI algorithms play a pivotal role in managing complex energy grids through predictive load balancing and proactive fault detection, significantly reducing downtime and enhancing overall reliability. Financial institutions leverage AI to safeguard assets and customer data through sophisticated fraud detection systems, automated algorithmic trading platforms, and enhanced risk management practices, which collectively ensure financial stability and consumer confidence.

Healthcare is another sector profoundly transformed by AI. With its capacity for precise data analytics and predictive modeling, AI supports critical diagnostic processes, personalized patient care, and efficient management of patient records. Particularly in scenarios such as epidemiological forecasting, AI-driven analytics can effectively predict outbreaks, thus enabling preemptive measures to safeguard public health.

In transportation systems, AI applications significantly improve operational safety and efficiency through traffic flow optimization and autonomous vehicle management, reducing congestion and lowering the likelihood of accidents. Additionally, critical manufacturing and chemical industries utilize AI technologies primarily for predictive maintenance and hazard identification, thereby minimizing the risk of industrial accidents and improving compliance with stringent safety regulations.

Moreover, sectors such as defense and emergency services heavily depend on AI for mission-critical applications. Autonomous drone operations, real-time threat assessment, and predictive policing are some prominent examples of how AI enhances the effectiveness and responsiveness of defense and public safety initiatives. Likewise, AI-powered surveillance and threat detection technologies greatly enhance the security of communications and government facilities, rapidly identifying and mitigating potential threats.

Table 1: U.S. Critical Infrastructure

U.S. Critical Infrastructure			
Chemical	**Dams**	**Financial Services**	**Information Technology**
Commercial facilities	Defense Industrial Base	Food and Agriculture	Nuclear Reactors, Materials, and Waste
Communications	Emergency Services	Government Facilities	Transportation Systems
Critical Manufacturing	Energy	Health Care and Public Health	Water and Wastewater

Table 2 illustrates the many uses of AI within critical infrastructure sectors where the AI often increases the resilience and security of the critical infrastructure components through both physical and cyber threats.

Table 2: U.S. Critical Infrastructure Sectors Leveraging Artificial Intelligence

Sector	**Typical AI Common Uses in the Sector**
Chemical	Predictive maintenance, process optimization/efficiency and hazard identification for equipment failure.
Dams	Monitoring dam safety through real-time data analysis of structural integrity, water levels, and potential overflow predictions for risk management/disaster prevention.
Financial Services	Fraud detection, risk management, algorithmic trading for market status, and customer service automation.
Information Technology	Intrusion detection systems and automated threat intelligence for near real time cyber-attack predictions of potential attack patterns.
Commercial Facilities	Security surveillance, energy management, and operational logistics throughout facilities expediting detection.
Defense Industrial Base	Autonomous drones, command & control (C2) systems, predictive maintenance for equipment, and cybersecurity.
Food and Agriculture	Optimize crop management and monitoring through precision agriculture techniques related to crop health, optimization of water/fertilizer and crop yield prediction.
Nuclear Reactors, Materials, and Waste	Monitoring and maintaining safety standards in nuclear facilities for optimization, failure detection, and safety compliance.
Communications	Security surveillance of unusual activities and energy management optimization.
Emergency Services	Enhance dispatch systems, predictive policing of crime patterns, and emergency response strategies for optimization of resource allocation.
Government Facilities	Security enhancements through facial recognition technologies and threat detection systems to enhance efficiency.
Transportation Systems	Improve traffic management, autonomous vehicle navigation, and predictive maintenance for vehicles and infrastructure for optimization while reducing potential traffic congestion areas.
Critical Manufacturing	Predictive maintenance, quality control, and supply chain optimization in manufacturing where component failure and efficiency are monitored to reduce downtime.
Energy	Energy distribution and integration of renewable energy sources into the grid leverages AI to simulate load forecasting for supply and demand usage.
Health Care and Public Health	Diagnostic procedures, patient management, and epidemiological research are enhanced with increased diagnostic accuracy, optimization of treatment and forecasting potential of disease outbreaks.

Sector	Typical AI Common Uses in the Sector
Water and Wastewater	Water quality monitoring, leak detection, and predictive maintenance of water infrastructure often use AI to analyze failure trends and optimize the safe production and efficient use of water.

The expansive use of AI within U.S. critical infrastructure sectors has proven indispensable for maintaining operational continuity and securing national interests. As AI-driven capabilities continue to mature, the potential for further enhancements in efficiency, safety, and cybersecurity remains significant. However, with these advancements comes the responsibility of ensuring robust protection mechanisms to counter evolving threats. Establishing clear regulatory standards and frameworks will be essential to manage these growing complexities effectively. Ultimately, embracing AI with structured risk management frameworks will position critical infrastructure sectors to better anticipate, withstand, and respond to future challenges.

4. Research Methodology

The research methodology employed in this study explicitly focuses on leveraging two critical frameworks: the MIT AI Risk Repository and the MITRE ATLAS (MIT, 2025; MITRE, 2024). This targeted methodology enables a systematic and comprehensive assessment of security vulnerabilities associated with AI systems within critical infrastructure sectors. This research utilized the theory-practice gap to bridge the common issue of technology development to applying theoretical insights directly to practical problems. This process allows for a better understanding of the complexities within these systems and fosters the development of actionable solutions (Aven, 2023; Bernard, 1999).

In the context of software development and artificial intelligence (AI), the theory-practice gap is particularly noticeable. Theoretical knowledge often guides the development of AI capabilities, but practical challenges emerge when these theories are applied in real-world coding and development. To address these challenges, developers engage in collaborative efforts, such as code refactoring, to incrementally improve the codebase. This process involves developing a code refactoring plan, writing automated tests, and making iterative changes to enhance the code's functionality and security (Bleher & Braun, 2023; Ouhbi, 2024). This research presents two research questions: Research Question 1: Can the MIT AI Risk Repository and the MITRE ATLAS improve AI security? Research Question 2: Can the MIT AI Risk Repository and the MITRE ATLAS be effective for the non-security AI engineer?

The MIT AI Risk Repository framework and the MITRE ATLAS framework directly pertain to vulnerabilities pertinent to critical infrastructure sectors—including energy, finance, healthcare, transportation, and defense. The MIT AI Risk Repository framework specifically focused on causal taxonomy (responsible entity: human or AI; intent: intentional or unintentional; occurrence stage: pre- or post-deployment) and domain taxonomy, offering a structured categorization to facilitate clear and actionable insights. The MITRE ATLAS framework was utilized to provide a comprehensive review of adversarial tactics, techniques, and procedures (TTPs) specifically targeting AI systems and provided detailed insights into real-world incidents, highlighting critical threats such as data poisoning and adversarial manipulations. The combination created a robust linkage between known adversarial actions and potential vulnerabilities (MIT, 2025; MITRE, 2024). To validate and enhance the practical applicability of these frameworks, a systematic literature review was performed using Cornell's ARXIV database, spanning articles published between 2020 and 2024. The literature search specifically targeted three significant AI security concerns: deepfake technology, AI system poisoning, and supply chain security. This review identified a total of 269 relevant academic articles, including 189 addressing deepfake security concerns, 28 focusing on AI system poisoning, and 21 related to AI supply chain vulnerabilities (Cornell University, n.d.). Articles were analyzed to identify existing vulnerabilities, adversarial strategies, and recommended mitigation measures, all assessed in alignment with the capabilities provided by the MITRE ATLAS and MIT AI Risk Repository frameworks.

This methodology, combining structured framework analysis and a systematic literature review, ensures recommendations that are both theoretically sound and practically implementable, effectively addressing evolving threats facing AI-dependent critical infrastructure. While the limitations of this research were restricted to deepfake security concerns, AI system poisoning, and AI supply chain vulnerabilities, these three vulnerabilities constitute some of the most critical vulnerabilities that currently exist with AI systems.

5. Framing the Problem

The MIT AI Risk Repository is the first attempt to address the urgent need for a comprehensive policy along with the technical response necessary to secure AI systems against a wide range of attack vectors that address the

wide reaching national security and societal impacts. (Comiter, 2019; MIT, 2025) This AI Risk Repository is a living database with 1000+ risks with 56 frameworks and classifications of AI risks. It is accessible and updatable through the MIT AI Risk Repository website. The causal taxonomy categorizes the how, when and why of these risks while the domain taxonomy lists the 7 major domains and 23 subdomains. The causal taxonomy classifies risks based on their causal factors. Some of these causal factors are whether the responsible entity was human or AI, whether it was intentional or unintentional, and whether the occurrence happened pre- or post-deployment. (MIT, 2025) The domain taxonomy considers the risk across the seven broad sub-domains and 23 subdomains to address possible outcomes from the risk.

Causal Taxonomy of AI Risks

From a provider and policymaker standpoint, AI vulnerabilities and compliance, negative outcomes could come from adversaries, unintentional acts, data with bias or error, and software bugs and configurations installed in the system. Adversaries could manipulate the system through changes in the algorithms or parameters. Human error by unintentional users could cause negative impacts to the system affecting the performance of the system. Bias or error introduced to the data set used for training could cause the AI system to draw incorrect conclusions. Software bugs and configurations installed in the system could also cause the system to behave in less than adequate ways. (Comiter, 2019) The MIT AI Risk Repository considers Causal Taxonomy of AI Risks from the aspects of category, level, and description of risk presentation. (MIT, 2025) Figure 1 illustrates the Causal Taxonomy of AI Risks.

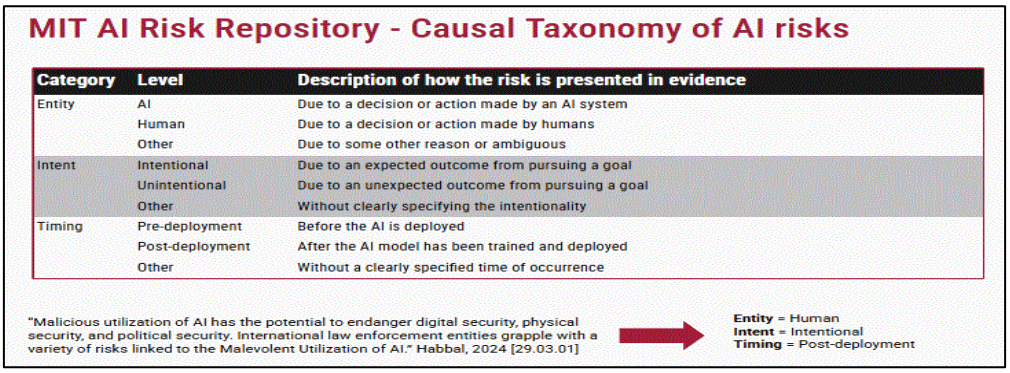

Figure 1: MIT AI Risk Repository - Causal Taxonomy of AI Risks

Domain Taxonomy of AI Risks

While causal taxonomy points to the actor, the domain taxonomy builds on the subdomains affected by the causation to describe the outputs of AI Risks. These subdomains include discrimination & toxicity, privacy & security, misinformation, malicious actors & misuse, human computer interaction (HCI), socioeconomic & environmental harms, and AI system safety, failures, and limitations as shown in figure 2. (MIT, 2025)

Figure 2: MIT AI Risk Repository - Domain Taxonomy of AI Risks

These subdomains have direct impacts on societal systems, military, law enforcement, and commercial providers of critical infrastructure. With many critical infrastructure sectors leveraging AI as shown in table 2, there is a greater need for AI security compliance, regulatory recommendations and policy/enforcement from government entities. (MIT, 2025) AI security compliance encourages like-in-kind compliance standards within critical infrastructure sectors so that stakeholders adopt best practices during AI system deployments, system updates and develop attack response strategies. Regulatory recommendations and policy/enforcement helps ensure compliance by public and private sectors providers to ensure mandatory compliance and prevent high-risk AI uses from impacting societal ecosystems consuming AI-enabled critical infrastructure. Policy and enforcement set the standard for planning, deployment, post-deployment, and security issue mitigation with compliance leveraged at each phase. (Comiter, 2019)

6. Adversarial Perspective with MITRE ATLAS

The MITRE Adversarial Threat Landscape for AI Systems (ATLAS™) was released in 2021. It was adapted from the MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) framework which was first released in 2013 and modeled for AI Systems. (MITRE, 2024a; MITRE, 2024b). MITRE ATLAS's framework is a constantly updated and serves as a living knowledge base that documents real-world attacks against AI systems where it provides insights from these attacks. These insights often providing attribution and adversary TTPs captured by AI red teams and security groups. Understanding TTP helps organizations better detect, understand, and mitigate these threats. The MITRE ATT&CK Framework developed an anonymous incident reporting portal where users can submit anonymous details; this capability has been extended to directly support the MITRE ATLAS AI Incidents Reporting Framework through a separate portal where users can submit and receive anonymized community AI incident information. (MITRE, 2024c)

Following the MITRE ATT&CK framework, MITRE ATLAS has 14 different categories with multiple techniques under each category for assessing the security of an AI system from adversarial threats. These 14 adversarial threats categories are: reconnaissance, resource development, initial access, ML model access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, collection, ML attack staging, exfiltration, and impact. Standard practices like protection from unauthorized access, ensuring data integrity, and AI hardening can be approached through these 14 categories to ensure AI system security posture and resilience. Some of the AI-specific threats are data poisoning where structured strategies are illustrated with how to mitigate their occurrence. The MITRE ATLAS Matrix Framework can also be used to help meet regulatory compliance by helping organizations stay informed on emerging standards and controls that are essential for the security of AI systems. The MITRE ATLAS Matrix Portal is shown in figure 3 (MITRE, 2024a).

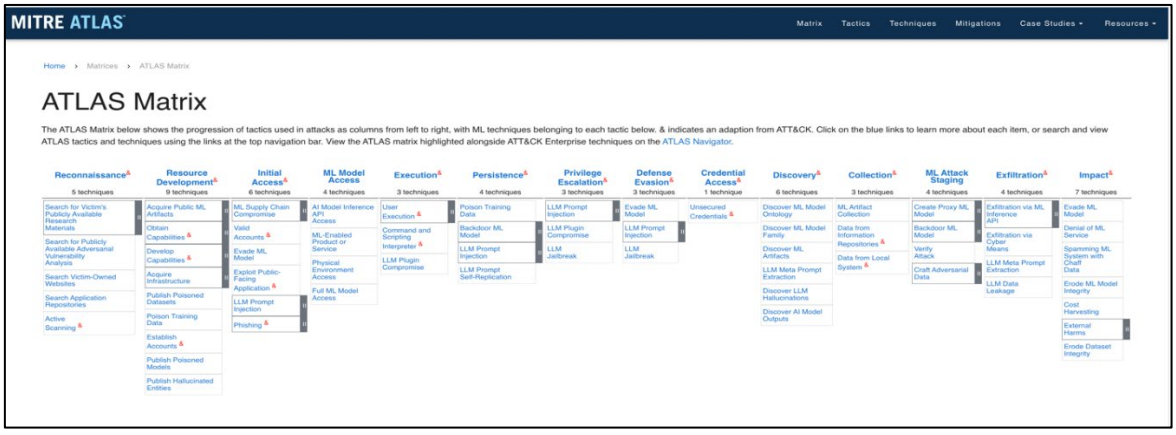

Figure 3: MITRE ATLAS Matrix Portal

7. Using the MIT AI Risk Repository and the MITRE AI ATLAS on a Typical Example: AI System Poisoning

The AI System Poisoning risk exemplar was selected for the exemplar because of the possibility of poisoning the system with minimal alterations (0.01% of a dataset) and least cost to an adversary ($60 USD). This is extremely significant when you consider how easy it is to contaminate an AI system. Searching through the 1,520 entries MIT AI Risk Repository using the "AI Risk Database v2" tab querying the risk subcategory for "poisoning" identified three entries shown in figure 4. The high-level causal taxonomy indicates the entity is "human" with

two “intentional” and one “other” intent reason as being generally introduced during “pre-deployment” and affecting the sub-domains of “AI system security vulnerabilities and attack” and “exposure to toxic content.”

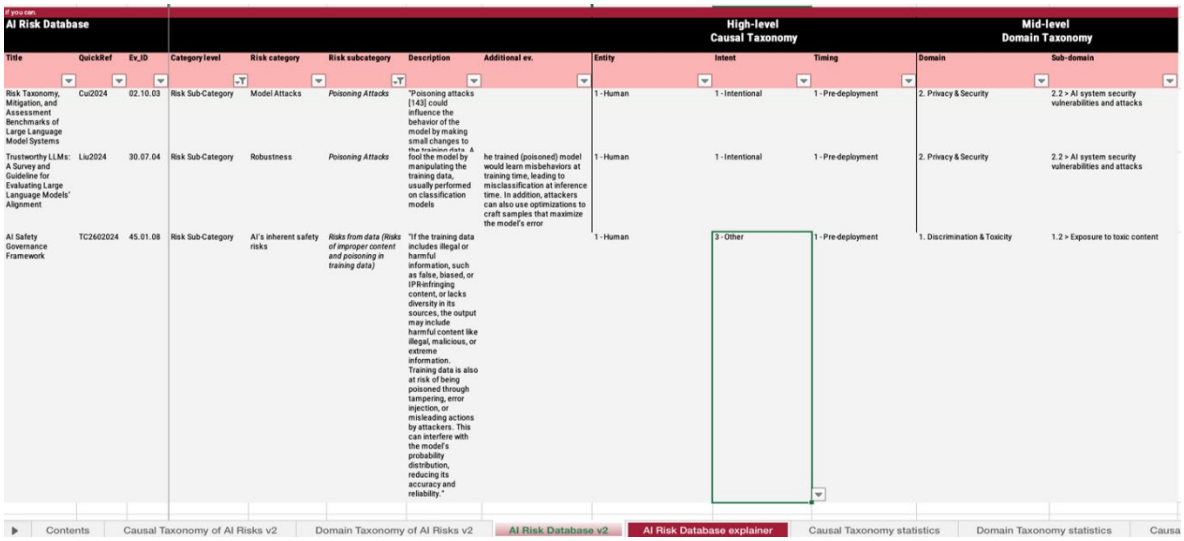

Figure 4: MIT AI ATLAS Query on Poisoning

Using the MITRE AI ATLAS Navigator link shown on the upper portion of the MITRE ATLAS Matrix Portal takes the user to a Github-hosted search environment shown in figure 5. Using the search icon at the top of the Navigator tool and entering “poisoning” yields 5 techniques: Adversary-in-the-Middle: ARP Cache Poisoning, Adversary-in-the-Middle: LLMNR/NBT-NS Poisoning and SMB Relay, Network Sniffing, Poison Training Data and Stage Capabilities: SEO Poisoning as commonly used adversary attack TTP with each have a detailed description of how the attack is initiated. No specific adversaries have been assigned for this set of exploits at this time but under the software subsection, typical Advanced Persistent Threat (APT) groups that have used this exploit in other MITRE ATT&CK framework were Ember Bear, APT28 and the Lazarus Group. Mitigation techniques are through the “Validate ML Model” with three approaches: ML Supply Chain Compromise: Model, Backdoor ML Model: Poison ML Model, and Backdoor ML Model: Inject Payload. These Validate ML Model approaches provide a case study detailing how the attack occurred (MITRE, 2025).

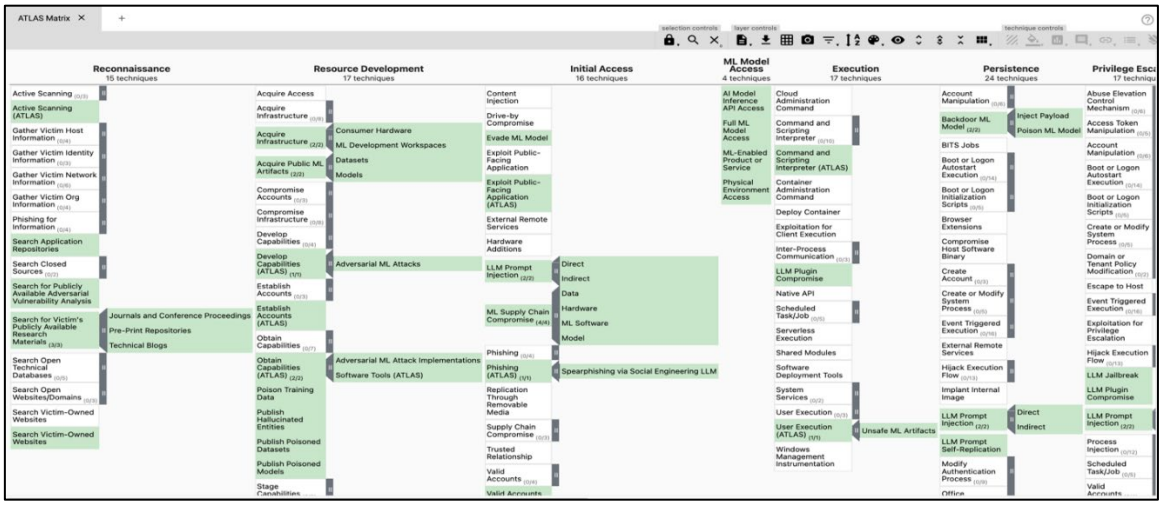

Figure 5: MITRE ATLAS Matrix Portal: ATLAS Navigator

This exemplar illustrates the approach and ease of a non-security person could research potential threats, vulnerabilities, impacts, and mitigations used in an AI System Poisoning attack vector that could lead to better protecting their AI system. Having this ease of use could help all AI engineers in improving the security posture of their capability.

8. Conclusion

This research emphasizes the crucial need for a proactive and comprehensive cybersecurity approach as AI integrates into critical infrastructure, advocating for advanced detection systems, robust security protocols, and international collaboration to establish universal standards, utilizing resources like the MIT AI Risk Repository and MITRE ATLAS. It accomplishes this by pointing out AI's ever-growing role and cyber risks by showing how AI brings expanding capabilities but also introduces vulnerabilities to sophisticated cyberattacks. Only through proactive and comprehensive security at all stages of development and deployment can these cyberattacks be mitigated. Utilizing the MIT AI Risk Repository and the MITRE ATLAS frameworks lead to a structured method for identifying and mitigating AI-specific threats. While these two frameworks are nascent and continually evolving, they provide a useful tool for non-security personnel with an approach to identify potential threats, vulnerabilities, impacts and mitigations. These tools also suggest a benchmark for where global cybersecurity practices need to be focused for critical infrastructure AI-related system. Additionally, future research should include international collaboration and universal standards where research can be expanded to further these frameworks and research to improve the resilience of critical systems against a wide range of threats.

***Disclaimer**: All statements of fact, opinion, or analysis expressed are those of the author. The views and opinions expressed herein by the author do not represent the official policies or positions of the United States (U.S.) Department of Defense (DoD), U.S. Navy, or other agencies or departments of the U.S. government and are solely representative of the views of the author. This does not constitute an official release of DoD or Navy information.*

References

Aven, T., 2023. On the gap between theory and practice in defining and understanding risk. *Safety science*, *168*, p.106325.

Bernstein, D.S., 1999. On bridging the theory/practice gap. *IEEE Control Systems Magazine*, *19*(6), pp.64-70.

Biden, J. R., 2021. *The U.S. President's Executive Order (EO) 14017, America's Supply Chain: A Year of Action and Progress*, The White House, Washington, DC, [https://www.whitehouse.gov/wp-content/uploads/2022/02/Capstone-Report-Biden.pdf].

Bleher, H. and Braun, M., 2023. Reflections on putting AI ethics into practice: how three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics*, *29*(3), p.21.

Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K. and Tramèr, F., 2024, May. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 407-425). IEEE, [https://arxiv.org/abs/2302.10149].

Comiter, M., *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It (Belfer Center Science and International Affairs, Harvard Kennedy School, 2019)*, [https://www.belfercenter.org/publication/AttackingAI]

Cornell University, n.d., arXiv Open-Access Database – Computer Science Sub-Category, [https://arxiv.org/search/cs] (Accessed: November 08, 2024).

Gnad, D., Gotthard, M., Krautter, J., Kritikakou, A., Meyers, V., Rech, P., Condia, J.E.R., Ruospo, A., Sanchez, E., Dos Santos, F.F. and Sentieys, O., 2024, May. Reliability and Security of AI Hardware. In *2024 IEEE European Test Symposium (ETS)* (pp. 1-10). IEEE.

Massachusetts Institute of Technology (MIT), 2025, MIT AI Risk Repository, https://airisk.mit.edu/].

Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A. and Malik, H., 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 53(4), pp.3974-4026, [https://arxiv.org/pdf/2103.00484].

Murphy, H. (2023) 'Deepfakes make banks keep it real: AI-enhanced impersonators can be detected with new monitoring technology', *The Financial Times*, September 20, 2023, Available at: [https://www.ft.com/content/6ca90b12-3ee6-409c-968d-1cffe29ee973] (Accessed: November 08, 2024).

MITRE, 2024a, MITRE ATLAS Framework: Navigate threats to AI systems through real-world insights, [https://atlas.mitre.org/].

MITRE, 2024b, MITRE ATT&CK Framework, [https://attack.mitre.org/].

MITRE, 2024c, MITRE ATLAS AI Incidents: Submit and receive anonymized community AI incident, [https://ai-incidents.mitre.org/].

MITRE, 2025, ATLAS Navigator, [https://mitre-atlas.github.io/atlas-navigator/#layerURL=https://raw.githubusercontent.com/mitre-atlas/atlas-navigator-data/main/dist/default-navigator-layers/atlas_layer_matrix.json]

NSA, 2022a, NSA, CISA, ODNI Release Software Supply Chain Guidance for Developers. [https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3146465/nsa-cisa-odni-release-software-supply-chain-guidance-for-developers/].

NSA, 2022b, ESF Partners, NSA and CISA Release Software Supply Chain Guidance for Customers, [https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3221208/esf-partners-nsa-and-cisa-release-software-supply-chain-guidance-for-customers/].

NSA, 2022c, Securing the Software Supply Chain: Recommended Practices for Developers, [https://media.defense.gov/2022/Sep/01/2003068942/-1/-1/0/ESF_SECURING_THE_SOFTWARE_SUPPLY_CHAIN_DEVELOPERS.PDF]

Ouhbi, S., 2024, April. Bridging the Theory-Practice Gap in a Maintenance Programming Course: An Experience Report. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training* (pp. 359-367).

Tong, Z., (2023), 'Google AI expert warns of 'data poisoning' as Chinese scientists work to ward off emerging threat', *South China Morning Post*, July 11. Available at: [https://www.scmp.com/news/china/science/article/3227213/google-ai-expert-warns-data-poisoning-chinese-scientists-work-ward-emerging-threat] (Accessed: January 04, 2025).