Adversarial AI threats on digitization of Nuclear Power Plants

Kishor Datta Gupta 1, Sajedul Talukder 2, and Syed Bahauddin Alam 2

1Clark Atlanta University
2Affiliation not available

October 31, 2023

Abstract

The future digitalization of Nuclear power plants (NPP) involves uses of sensors data in digitalize formats and analysis by AI based techniques. The planned and existing network architectures used by U.S. nuclear facilities, particularly in high-security zones near the reactor cores, typically rely on a private intranet that connects multiple computers in a peer-to-peer (P2P) or similar network configuration. These intranets are considered highly secure due to their isolation from the outside world (Internet) and the implementation of data diodes to regulate one-way data flow. Additionally, these facilities employ various security measures, including physical security, authorized access, pre-screened employees, antivirus software, and supply chain verification. While these “air gap” systems and their security frameworks are effective in protecting against traditional threats like malware, ransomware, trojans, and intrusion detection, they may overlook a growing vulnerability related to AI models within these air gap systems. AI models or decision systems utilized within nuclear facilities collect sensor data through the secure network and make critical decisions. However, despite the network’s robust security measures, the threat to AI models posed by adversarial attacks such as troj-AI, evasion-based attacks, backdoor attacks, and pre-trained poisoned attacks is often ignored by conventional virus scanners. These attacks exploit the vulnerabilities of AI models and can be difficult to detect without domain-specific knowledge related to the model data. As the security of nuclear power plants is paramount, it is crucial that we proactively scan and monitor AI models used in different sectors of these facilities. Unfortunately, there is currently no established framework to monitor and scan the behaviors and architectures of AI models, which poses a significant vulnerability for nuclear power plants. To ensure the comprehensive security of nuclear facilities, it is necessary to address this gap by developing specialized frameworks and mechanisms to monitor and assess the security of AI models. These frameworks should be capable of detecting and mitigating adversarial attacks targeting AI models, providing an additional layer of protection alongside existing security measures. By proactively addressing this emerging threat, we can enhance the overall security posture of nuclear power plants and better safeguard against potential risks.
Adversarial AI threats on digitization of Nuclear Power Plants

Kishor Datta Gupta,* Sajedul Talukder,† Syed Bahauddin Alam ‡∗

* Computer Science, Cyber Physical System, Clark Atlanta University, Atlanta, USA
† Computer Science, Southern Illinois University, Carbondale, IL, USA
‡ Nuclear, Plasma & Radiological Engineering, University of Illinois at Urbana-Champaign, IL, USA
∗ National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL, USA

INTRODUCTION

The future digitalization of Nuclear power plants (NPP) involves uses of sensors data in digitalize formats and analysis by AI based techniques. The planned and existing network architectures used by U.S. nuclear facilities, particularly in high-security zones near the reactor cores, typically rely on a private intranet that connects multiple computers in a peer-to-peer (P2P) or similar network configuration. These intranets are considered highly secure due to their isolation from the outside world (Internet) and the implementation of data diodes to regulate one-way data flow. Additionally, these facilities employ various security measures, including physical security, authorized access, pre-screened employees, antivirus software, and supply chain verification. While these "air gap" systems and their security frameworks are effective in protecting against traditional threats like malware, ransomware, trojans, and intrusion detection, they may overlook a growing vulnerability related to AI models within these air gap systems. AI models or decision systems utilized within nuclear facilities collect sensor data through the secure network and make critical decisions. However, despite the network’s robust security measures, the threat to AI models posed by adversarial attacks such as troj-AI, evasion-based attacks, backdoor attacks, and pre-trained poisoned attacks is often ignored by conventional virus scanners. These attacks exploit the vulnerabilities of AI models and can be difficult to detect without domain-specific knowledge related to the model data. As the security of nuclear power plants is paramount, it is crucial that we proactively scan and monitor AI models used in different sectors of these facilities. Unfortunately, there is currently no established framework to monitor and scan the behaviors and architectures of AI models, which poses a significant vulnerability for nuclear power plants. To ensure the comprehensive security of nuclear facilities, it is necessary to address this gap by developing specialized frameworks and mechanisms to monitor and assess the security of AI models. These frameworks should be capable of detecting and mitigating adversarial attacks targeting AI models, providing an additional layer of protection alongside existing security measures. By proactively addressing this emerging threat, we can enhance the overall security posture of nuclear power plants and better safeguard against potential risks.

PROBLEM DISCUSSION

We present our assumptions that (i) The NPP employees, supply chain, entry security routines, etc., may have been compromised and hence can provide adversarial input into one of the sensors inside the high-security zone; (ii) The AI systems inside Air-gap may have a trojan trigger (Installed by the malicious agent in the time of deployment or development) to generate wrong decision (iii) There may have unintentional bias inside AI systems that may be making wrong decisions. (v) The adversarial attack has the capability of disrupting safety related systems (as they depend on sensors data), compromising not only the reactor itself but also environmental health and personnel safety. Under these assumptions, any type of contemporary anti-virus and anti-malware technique will be insufficient to prevent the installation and misbehaving of such sophisticated AI systems under adversarial attacks. This is where our discussion in this paper is crucial to conceptualize, prevent and mitigate the impact of such highly posed adversarial threats.

Adversarial AI Threats

Three basic types of adversarial attacks are: data poisoning, evasion-based, and model manipulation-based attacks. Assuming that our models are prepared in a secure manner, we can disregard data poisoning attacks. Instead, we will focus on evasion-based and model manipulation-based attacks, which are possible when a target ML model is deployed in an air-gapped system. Evasion-based Adversarial attacks (AAs) employ various cunning techniques to manipulate input data by adding traits or noises, posing a significant threat to the reliability of deep learning models. The reasons behind the success of AAs remain inconclusive. Szegedy et al. (2014)[1] attributed their success to non-linearity, while Goodfellow et al. (2014)[2] argued that it is due to excessive linearity. Tanay and Griffin (2016) [3] introduced the tilted boundary theory, suggesting that it is impossible to perfectly fit a model, thereby allowing for the existence of AAs [4]. MIT researchers proposed that adversarial features are not merely noise, but rather data that cannot be accurately classified due to limitations in human sensors. However, this viewpoint is challenged by other researchers [5]. A variation of the model manipulation-based attack is known as Troj-AI. If a deep learning network architecture is modified by an adversarial agent after its training is completed, then that deep learning network has a Troj-AI [6] [4]. In Figure 1, we illustrate how a trojan-trigger in the network can change the output results by manipulating the network. It first gathers information about what the input data should be in order to be classified as a certain class. Based on that, it changes the architectural weights. Another example of Troj-AI is the Backdoor system [7], which is illustrated in Figure 2.
Fig. 1. Trojan AI[8]

Fig. 2. Backdoor attack[7]

Adversarial AI Threats in Perspective of Nuclear Power Plant

The threats discussed (illustrated in figure 3 in the previous subsections) are also applicable in an airgap system under the following conditions:

- Sensors can be physically altered by insiders: An adversarial patch attack[9] can be used to modify the sensors, potentially enabling a trigger for a trojan within the AI model.

- Sensor data communication path can be attacked: The path through which sensor data is communicated to the AI model within the intranetwork can be compromised or modified by an attacker.

- Installation medium of AI model can be tampered with: The AI model used in airgap systems is typically deployed using a digital medium, which can be subject to attacks that manipulate the model during installation.

Fig. 3. Vulnerable points to initial Adversarial AI attacks

- Closed source model Extractions: Even if these models are considered closed-source, bitflip attacks can potentially recover the model architecture[10]. This recovered information can then be used to trigger a trojan within the airgap system by exploiting sensor data.

In summary, these conditions demonstrate how the aforementioned threats can still pose a risk to an airgap system, highlighting the need for robust security measures to mitigate such vulnerabilities.

Limitations of current defenses

Current defense techniques in cybersecurity research primarily focus on software security and network security, often overlooking the security of the AI model itself. For example, there has been limited research on identifying perturbation attacks or detecting trojan triggers within AI models. While measures such as firewalls, intrusion detection systems, and access controls contribute to protecting the network and preventing unauthorized access, they do not provide comprehensive protection against adversarial attacks targeted at AI models. It is particularly challenging to identify trojan triggers in black box models, which are commonly employed in nuclear power plant (NPP) systems. As example currently employed data-driven technology in NPP is MSET1 (currently in use at EBR-II reactor and Crystal River 3 nuclear power station), which is a software system for real-time process monitoring. MSET-based techniques are not immune to adversarial attacks from a cybersecurity perspective. Researchers have already demonstrated the formulation of adversarial attacks on MSET models[?, ?]. Moreover, one notable characteristic of adversarial attacks is their transferability between models. This means that adversarial inputs generated for neural network-based models can also be effective against MSET-based models. The use of historical data in MSET models introduces potential vulnerabilities to poison-based attacks. Adversaries could inject malicious data into the training dataset, leading to compromised model performance and erroneous estimations. Additionally, the increasing randomness associated with sensor readings creates an opportunity for attackers to insert a trojan trigger within the MSET model. By exploiting this randomness, an adversary could trigger specific behaviors in the system that deviate from expected norms. Moreover, the current defense techniques in airgap models have predominantly emphasized software and network robustness, thereby neglecting the specific security concerns associated with AI models. In summary, there is

1https://www.anl.gov/nse/ai-ml/mset
a clear requirement for extensive research to develop effective defense techniques to identify and mitigate perturbation attacks, detect trojan triggers within black box models, and provide comprehensive protection against adversarial attacks. By addressing these gaps in research, we can enhance the overall security posture and robustness of AI models, particularly in airgap systems and critical infrastructure like NPPs.

**Proposed strategies**

To enhance the security of AI models in airgap systems, several areas of research should be explored:

- **Model introspection and explainability**: Research on techniques that enable better understanding and transparency of AI model behavior. This includes methods for identifying and interpreting the internal representations and decision-making processes of the models, which can aid in identifying trojan triggers and potential vulnerabilities.

- **Threat modeling and risk assessment**: Comprehensive threat modeling and risk assessments specific to airgap systems and NPP environments. This helps identify potential attack vectors, prioritize security measures, and allocate resources effectively to address the most critical risks.

- **Adversarial attack protection**: Developing techniques to improve the robustness of AI models against perturbation attacks, including adversarial training, defensive distillation, and input sanitization methods. These approaches can detect and mitigate the impact of trojan triggers within the models.

- **Adversarial input detection**: Implement a system to identify adversarial perturbations in sensors input. As detecting adversarial input will help to identify insiders in the airgap system, instead of protecting AI model from perturbation, identifying adversarial input detection is more important.

- **Identify Trojan inside AI**: Identify troj-AI by scanning the architecture or other AI-specific v&V techniques. Troj-AI can be intentional and non-intentional too. The AI-model developer may or may not compromise the AI model, but still a troj-AI can exist. Establishing rigorous procedures to validate and verify the integrity and security of AI models before their deployment in airgap systems. This involves extensive testing, code review, and formal verification techniques to detect any potential trojan triggers or vulnerabilities.

By emphasizing these research directions, the security of AI models in airgap systems, including NPPs, can be significantly enhanced.

**CONCLUSION**

In conclusion, the increasing digitalization of nuclear power plants (NPPs) introduces new security challenges, particularly concerning the protection of AI models within airgap systems. While traditional defense techniques focus on software and network security, they often overlook the security of AI models, leaving them vulnerable to adversarial attacks. The potential consequences of such attacks within NPPs, including compromised safety systems and personnel safety, emphasize the need for proactive security measures. To address this gap, specialized frameworks and mechanisms are required to monitor and assess the security of AI models in airgap systems. These frameworks should be capable of detecting and mitigating adversarial attacks, such as troj-AI and evasion-based attacks, providing an additional layer of protection alongside existing security measures. Research areas like adversarial robustness, model introspection, validation and verification, threat modeling, hardware-assisted security, and collaboration are essential for enhancing AI model security in airgap systems. Investing in research and development in these areas can improve the overall security posture of nuclear power plants. It is crucial to stay ahead of emerging threats, prioritize the security of AI models, and foster collaboration among researchers, industry experts, and regulatory bodies. Through these efforts, we can better safeguard critical infrastructure, protect against potential risks, and ensure nuclear power plants’ safe and reliable operation.

**REFERENCES**