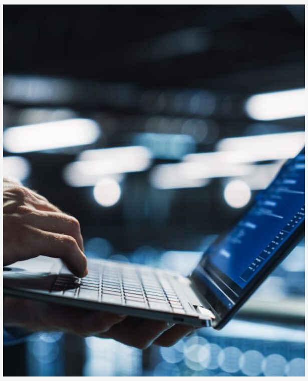
OSLER

Emerging AI security risks and considerations: key takeaways from the NIST adversarial machine learning report

MAY 10, 2024 6 MIN READ



Related Expertise

- Artificial Intelligence
- Commercial Technology Transactions
- Cybersecurity and Security Incident Response
- Disputes
- Emerging and High Growth Companies
- Privacy and Data Management
- Risk Management and Crisis Response
- Technology

Authors: Sam Ip, Naomi Chernos, Joseph Ierullo

As AI systems become increasingly prevalent and their integration into organizations become more entrenched, businesses face novel and evolving security and privacy risks. These risks stem largely from systems powered by machine learning models that are vulnerable to a class of security risks known as adversarial machine learning (AML), which target AI systems, including machine learning models and data used to train and serve such models. These risks will require innovative approaches as organizations learn to adapt and respond to such threats.

To help organizations familiarize themselves with the types of attacks and risks they might expect, as well as approaches to help mitigate them, earlier this year, the U.S. National Institute of Standards and Technology (NIST) published a comprehensive report intended to act as a cybersecurity guide for organizations that develop and oversee the governance of AI systems. The report is entitled "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" [PDF]. This report, which forms part of NIST's effort to support the development of trustworthy AI, helps give practicality to NIST's AI Risk Management Framework. Given the global influence of NIST's standards and guidance, Canadian organizations can gain insights into the direction and guidance that may develop from NIST's leading framework.

Specifically, the report introduces a taxonomy of different major classes of attacks for two broad categories of AI systems: predictive AI, and generative AI, based on attacker goals and capabilities, while providing corresponding mitigations. The report also discusses remaining challenges in the field and key developments still to come to ensure a trustworthy AI ecosystem.

Types of attacks

The NIST report classifies potential attacks by capability and objective, as well as identifying the stage of the model learning process where these attacks may be mounted. The goal in this classification is to create a standard terminology to aid in developing consistent strategies for mitigation. These classes of attacks include:

- Evasion attacks: occur after an AI system is deployed and attempts to alter an input to circumvent intended model behaviour and change how the system responds to it, essentially seeking to generate adversarial examples, which can be misclassified during deployment. Evasion attacks can be launched whether or not the attacker has prior knowledge of the model architecture or training data. Examples would include adding markings to stop signs to make an autonomous vehicle misinterpret them.

- Poisoning attacks: occur in the training phase by introducing corrupt data. An example might include slipping frequent instances of inappropriate language into conversation records so that a chatbot would interpret these instances as common enough parlance to use in its own customer interactions.
- Privacy attacks: occur during deployment and are attempts to extract sensitive information about the AI or the data it was trained on. Examples of this type of attack might be where an adversary asks a chatbot numerous legitimate questions and uses the answers to reverse engineer the model to find its weak spots or guess at its sources.
- Abuse attacks/Prompt injection attacks: these attacks occur in generative AI models, based on the fact that these models are trained by scraping wide and often unverified swathes of data. These involve the insertion of incorrect information into a source, such as a compromised webpage or online document, which the AI then absorbs as legitimate.

Potential mitigation strategies

While the report offers suggestions for potential mitigation strategies that are specific to each class of attack, several of the mitigation principles address the key overarching risk that the data used to train AI systems may not be trustworthy and, in fact, relies on exposure to sources of publicly available data to properly develop, which opens it to the risk of negative interference. However, because the datasets used to train AI are too large for a person to successfully monitor and filter, mitigations are difficult to implement.

In general, the mitigation approach for the majority of the classes of attacks focuses on data purification and model sanitization at the high level, and requires frequent audits and testing. However, these sanitization techniques, the report makes clear, must be combined with cryptographic techniques that will verify the source and authenticity of AI systems, rather than relying on spotting errors in the data itself. One common way this may be implemented is by incorporating red teaming (establishing an internal group tasked with probing the system for weaknesses), which will be a critical component of the pre-launch testing and assessment process for all AI systems to pinpoint potential security flaws. This is consistent with the ISED Voluntary Code of Conduct for the Responsible Development and Management of Advanced Generative AI Systems, which prescribes the use of adversarial testing (i.e., red-teaming) to identify vulnerabilities.

Furthermore, while the report mentions potential mitigation strategies for AML attacks, these should be considered alongside traditional threats relating to poisoned supply-chain models, data breaches, and service vulnerabilities inherent in the machine learning systems themselves. These threats continue to jeopardize data confidentiality and the integrity of answers and predictions created by the model.

Looking ahead

This report highlights the diverse and advanced threat landscape for each class of AML attacks, and the mitigation strategies that may require a distinct skillset and technical expertise. To navigate threats associated with AML, organizations will need to prioritize AI governance and develop an AI governance strategy that includes an appropriate risk management framework. These strategies should emphasize risks related to the security and privacy of data, given the data-centric approach of machine-learning and t in some cases, where sensitive and personal information may be used to train machine learning models. For more information on the privacy considerations involved in AI models, see here. For organizations to successfully manage risk, an awareness approach must be engrained in an organization’s culture.

Additionally, it will be critical for organizations to ensure there are sufficient protections in their contracts, particularly when procuring AI systems from third party vendors. These contracts should specify appropriate uses and safeguards of AI systems and machine learning models, and contain appropriate safeguards and measures to ensure the models and their datasets are not compromised, as well as, if appropriate, requiring AI vendors to comply with evolving standards, demonstrate certifications, and facilitate audits of its practices. Proper risk allocation is essential in these contracts to address liability associated with risks. Attacks and risks may also be introduced through fourth parties throughout the supply-chain. Contractual strategies for ensuring safeguards, such as ensuring data vetting and monitoring are implemented at all stages in the cycle, will need to be considered.

In summary, the NIST report underscores the importance of vigilance and innovation in AI security and in creating a clarifying vocabulary and providing related risk mitigation approaches. The report provides guidance for organizations that seek to enhance their AI practices and support their journey in the deployment and adoption of trustworthy AI systems.