February
2025

ARTIFICIAL INTELLIGENCE AND THE INSIDER THREAT

JOB AID

CDSE Center for Development of Security Excellence

INTRODUCTION

What is Artificial Intelligence and How Can It be Used as an Insider Threat?

Artificial intelligence (AI) and machine learning (ML) allow for a wide range of applications. AI applications use digital technologies to create systems designed to perform tasks that require intelligence or modest reasoning. ML is the development of digital systems that improve performance on tasks over time through that machine's experience. The increasing availability of AI tools, such as ChatGPT, has made technology more accessible to the average person for tasks such as resume review, art creation, code development, and academic research.

Despite its wealth of positive, productive uses, AI can also pose a threat to national security through misuse during cyberattacks, disinformation campaigns, and the manipulation of critical infrastructure/systems. AI can be used by malicious actors to cause significant damage and disruption to national security and the organizations that defend it.

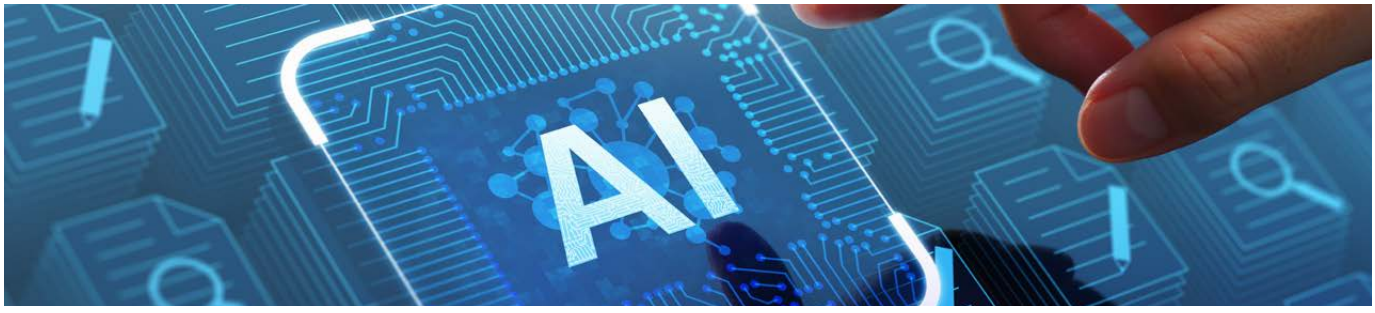
An "AI insider threat" in cybersecurity refers to a scenario in which an employee with access to an organization's AI systems intentionally misuses or manipulates them for malicious purposes, such as altering data, biasing algorithms, or stealing sensitive information, potentially causing significant damage to the organization. Examples include: a disgruntled employee modifying an AI-powered decision-making system to favor specific customers unfairly, a data scientist intentionally injecting bias into a machine learning model, or a malicious insider using AI tools to generate deepfakes for social engineering attacks.

What You Need to Know

- It is imperative that every organization acknowledges the potential harm that can be caused by insiders who misuse AI as a weapon for personal gain or to settle scores.
- In the hands of a motivated insider with even average technical proficiency, AI becomes a uniquely effective tool with which to penetrate an organization's complete security infrastructure.
- Continuous workforce screening is an irrefutable necessity in the arsenal of protection against AI-assisted insider threats.

Do you know how artificial intelligence affects you and your organization's operation? AI has the capability to provide near unlimited support to any application. It has the potential to benefit and promote further development in fields including but not limited to business, healthcare, communication, transportation, education, finance, and security.

Every organization must acknowledge and heed the potential harm that insiders who misuse AI as a weapon can cause. AI enables even a novice to target an organization's security infrastructure for any number of malicious purposes. While trusted people within your organization are its strongest resource, they are also its greatest vulnerability.



Learn more: Effective Cyber Security policies, assessments, and incident response plans can mitigate these risks and ensure DOD technology is delivered uncompromised. Click **here**.

UNDERSTANDING AI

AI as a Defensive Tool?

AI's advanced algorithms can filter through vast amounts of data to identify patterns or behaviors within a network, flagging potential risks before they escalate into serious issues. For example, AI can be trained to recognize the signs of potential data exfiltration or anomalous login activities, thus serving as a proactive measure against internal threats.

One of the most effective tools in this arsenal is User and Entity Behavior Analytics (UEBA). Sometimes referred to as User Behavior Analytics (UBA), these monitoring solutions utilize AI and ML algorithms to identify and halt behaviors that could signal an ongoing insider attack. For instance, a UEBA tool could instantly detect and disconnect a user who typically downloads only a small amount of data but suddenly starts downloading multiple gigabytes of data.

AI as a weapon?

The strengths that AI has can also be weaponized to assist insider threats. These models are becoming advanced enough to generate human-like text. This real-time communication can then be manipulated for malicious purposes, such as social engineering or data theft. While AI programs are built to make tasks more efficient, in the wrong hands or without proper governance, they can serve as powerful tools for malicious or negligent intentions.

Staff members unfamiliar or untrained in AI usage, especially regarding safeguards to confidential information, may become unintentional insider threats. Staff may unintentionally leak sensitive information to AI systems, which can then present a threat to organizations.

Learn more: Logging, monitoring, and auditing information system activities can lead to early discovery and mitigation of behaviors indicative of insider threats. Visit CDSE for additional training and resources: Click **here**.

TYPES OF AI-POWERED CYBERATTACKS

There are multiple types of cyberattacks that can be enhanced by AI and ML.

AI-driven social engineering attacks leverage AI algorithms to assist in the research, creative concepting, or execution of a **social engineering attack**. An AI-driven social engineering attack can:

- Identify an ideal target, including both the overall corporate target and a person who can serve as a gateway to the organization's information technology environment.
- Develop a persona and corresponding on-line presence to communicate with the attack target.
- Write personalized messages or create multimedia assets, such as audio recordings or video footage, to engage the target.

AI-driven phishing attacks use **generative AI** to create highly personalized and realistic emails, messages, voice communications, or social media to achieve a desired result. In advanced cases, AI can be used to automate the real-time communication used in phishing attacks.

Adversarial AI/ML is when an attacker aims to disrupt the performance or decrease the accuracy of AI/ML systems through manipulation or deliberate misinformation. Attackers use several adversarial AI/ML techniques that include:

- **Data collection and analysis**: Cybercrime groups, nation-state threat actors, and other parties are using AI tools to collect and analyze private data in greater detail than previously possible.
- **Evasion attacks**: Evasion attacks target an AI/ML model's input data. Attackers can leverage AI to enhance the efficiency of cyberattacks. AI-powered malware can adapt its behavior to evade detection by traditional cybersecurity measures, making it more challenging to detect and defeat threats.
- **Poisoning attacks**: Poisoning attacks target the AI/ML model training data, which is the information that the model uses to train the algorithm. In a poisoning attack, the adversary may inject fake or misleading information into the training dataset to compromise the model's accuracy or objectivity.
- **Model tampering/stealing**: Attackers can reverse-engineer AI applications by querying the model with questions and analyzing its responses. Once the model is understood, attackers can replicate its functionality or exploit its vulnerabilities. An adversary can also make unauthorized alterations to the model to compromise its ability to create accurate outputs.
- **Deepfakes**: Attackers can deceive others with AI-generated video, image, or audio files called deepfakes. Deepfakes are used to manipulate videos, audio recordings, and images for various malicious purposes, such as spreading misinformation, impersonating individuals, and conducting social engineering attacks. Deepfakes are usually part of a social engineering campaign.

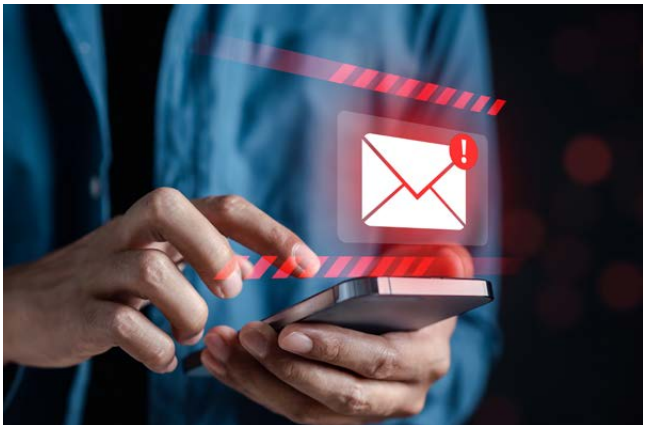
Malicious Generative Pre-Trained Transformer (GPT) is an AI model that responds to user prompts. A malicious GPT refers to an altered version of a GPT that produces harmful or deliberately misinformed outputs. A malicious GPT can generate attack vectors (such as malware) or supporting attack materials (such as fraudulent emails or fake online content) to advance an attack.

Ransomware attacks leverage AI to improve performance or automate some aspects of the attack path. AI can be leveraged to research targets, identify system vulnerabilities, encrypt data, and adapt and modify ransomware files over time, making them more difficult to detect with cybersecurity tools.

EXAMPLES OF AI INSIDER THREATS

June 5, 2023 – Artificial Intelligence used in Sextortion Schemes

Malicious actors use content manipulation technologies and services to exploit photos and videos — typically captured from an individual's social media account, open Internet, or requested from the victim — into sexually-themed images that appear true-to-life in likeness to a victim, then circulate them on social media, public forums, or pornographic websites. **Click here to read more.**



July 9, 2024 – U.S. Department of Justice Disrupts Russian Social Media Bot Farm

Russian actors created an AI-enhanced social media bot farm that spread disinformation in the U.S. and abroad. The social media bot farm used elements of AI to create fictitious social media profiles — often purporting to belong to individuals in the United States — which the operators then used to promote messages in support of Russian government objectives, according to according to unsealed affidavits. **Click here to read more.**

January 8, 2025 – Trader Arrested for Stealing Trade Secrets from Global Quantitative Trading Firm

A research developer and quantitative trader at a global, quantitative trading firm took advantage of their near complete access to the company's source code and stole valuable trade secrets to develop a competitor's source code. **Click here to read more.**

Learn more: Readers should further consult applicable and controlling laws, regulations, policies, and procedures. Visit CDSE for additional training and resources: Click **here**.

MITIGATING AI THREATS

Ensure data quality: Use diverse and unbiased training data.

Create threat intelligence: Develop a feed of AI-related threats to stay ahead of emerging threats.

Rotate encryption keys: Regularly update and rotate encryption keys used in AI.

Implement data and AI controls: Establish mechanisms to minimize privacy, security, and ethical risks.

Develop governance: Establish clear policies and structures for AI projects, including roles, responsibilities, and decision-making processes.

Create AI ethics principles: Create, implement, and operationalize AI ethics principles.

Oversee and monitor: Establish ongoing review and oversight of AI systems.

Enhance explainability: Make AI systems more explainable and interpretable.

Explore new techniques: Explore new risk-mitigating techniques, such as differential privacy and watermarking.

Learn more: CISA has developed a Roadmap for Artificial Intelligence, which is a whole-of-agency plan aligned with national AI strategy, to address our efforts to: promote the beneficial uses of AI to enhance cybersecurity capabilities, ensure AI systems are protected from cyber-based threats, and deter the malicious use of AI capabilities to threaten the critical infrastructure Americans rely on every day. Click **here**.

BEST PRACTICES FOR PREVENTING INSIDER THREATS

The risk to an organization's security and data can come from inside the organization, those who misuse AI tools and have access to sensitive data, causing harm through malicious, neglectful, or unintentional actions. Here are a few tips:

1. **Develop a Strong Security Culture**: Developing a culture of security is essential to preventing insider threats. This means providing regular security awareness training, monitoring employee behaviors, and enforcing security policies and procedures.
2. **Conduct Comprehensive Risk Assessments**: Conduct regular assessments of all vendors and third-party service providers.
3. **Conduct Background Checks**: Before hiring any new employees, it's important to conduct thorough background checks to ensure they have a clean record and no previous history of malicious activities.
4. **Use Least Privilege**: Limiting access to sensitive data and systems to only those employees who need it is another way to prevent insider threats. This practice is known as least privilege.
5. **Maintain Continuous Monitoring**: Keep a close eye on employee behavior and monitor their access to sensitive data and systems. This can help detect any abnormal or suspicious activity early on.
6. **Create a Strong Incident Response Plan**: Having a well-designed incident response plan is critical to responding quickly and effectively to any potential insider threats.
7. **Implement Information Sharing**: Collaboration and information sharing between private and public sectors can enhance the ability to detect and respond to sophisticated cyber threats.
8. **Implement Third-Party Risk Management**: Vendors and contractors should adhere to the same security standards as the contracting company itself.

By implementing these best practices, organizations can take proactive steps to prevent insider threats and minimize the risks associated with supply chain security. Remember, it only takes one bad actor to cause serious harm to your operations.

ADDITIONAL RESOURCES

Cyber Insider Threat – CDSE (eLearning Course)

Cybersecurity Attacks – CDSE (The Insider Threat Short)

Best Practices and Vulnerabilities for Privileged Accounts – CDSE (Webinar)

Potential Risk Indicators: Insider Threat – CDSE (Job Aid)

Insider Threat Indicators in User Activity Monitoring – CDSE (Job Aid)

Taking Culture Seriously – CDSE (Webinar)

Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats – Department of Homeland Security (Study)

Ensuring AI Is Used Responsibly – Department of Homeland Security

DHS Generative AI Public Sector Playbook – Department of Homeland Security

Impacts of Adversarial Use of Generative AI on Homeland Security – Department of Homeland Security (Science and Technology)

Insider Risk Management Program Office – U.S. Department of Commerce

Best Practices in Cyber Supply Chain Risk Management – National Institute of Standards and Technology (NIST)

Insider Threat Mitigation – Cybersecurity & Infrastructure Security Agency (CISA)