Courtesy of
Wheeler & Taylor Insurance

Defending AI Systems From Malicious Data Poisoning Attacks

As the use of artificial intelligence (AI) and machine learning (ML) continues to grow, businesses that utilize these technologies must also be aware of the attack methods cybercriminals use to target them. One such attack that hackers employ is data poisoning—when a malicious code is introduced into a dataset to compromise the performance of AI and ML systems.

Once installed, the unwelcome software manipulates training data to induce errors or biases, which can significantly decrease the reliability of these systems. Data corruption created by data poisoning can lead to critical errors that affect the accuracy and efficacy of AI system outputs, so businesses must ensure they have mechanisms to address this vulnerability.

This article provides more information on data poisoning attacks and tips to defend against them.

Data Poisoning Overview
By altering datasets during an AI's training phase, a hacker can compromise the integrity of the system's outputs, leading to errors, unintended results or biases. The attacks can also increase a system's vulnerability to additional cybersecurity issues by creating an access point for future intrusions.

There are several ways to carry out data poisoning, such as:

- Intentionally inputting incorrect or misleading information in the training dataset
- Modifying the existing dataset
- Deleting parts of existing datasets

Data poisoning attacks are generally classified based on their outcomes. Here are two common classifications:

1. **Targeted attacks** are when a malicious actor aims to influence the model's behavior in a specific situation. Targeted attacks generally do not impact the AI model's overall performance.
2. **Nontargeted attacks** occur when a cyber adversary attempts to manipulate the dataset to degrade the overall performance of the AI, thereby negatively affecting its predictive or decision-making abilities.

Threat Actors and Motivations
To address exposures, businesses must be aware of the different threats and the motivations behind these malicious actors. Examples of individuals or groups that may initiate data poisoning attacks include:

- **Malicious insiders**, including employees with access to data who may have grievances with an organization and seek revenge
- **External hackers** whose purpose is to exploit vulnerabilities to disrupt operations for financial gain
- **Nation-states** that seek to engage in cyberwarfare to undermine the technological advantages of their adversaries

Other parties involved in data poisoning may do so due to ideological beliefs. For instance, activists who look to increase privacy from AI may turn to data poisoning tactics to demonstrate flaws and vulnerabilities in AI to accomplish their objectives. Others may engage in these attacks to gain notoriety or to prove their capabilities. Whatever their motivations, businesses need to be aware of these potential infiltrations and take steps to mitigate their risks.

Examples of Data Poisoning Attacks
Malicious actors are discovering new ways to leverage data poisoning attacks. Strategies include:

- **Spam filter malfunctions**—A hacker can poison a dataset of an AI, allowing spam emails to bypass

filters and impact large numbers of employees and create vulnerabilities to other cyberattacks (e.g., phishing scams).

- **Network traffic misclassification**—A threat actor can poison a learning model's dataset to incorrectly label network traffic (e.g., web browsing and video streaming), leading to poor network performance.
- **Cybersecurity degradation**—Intrusion detection systems' datasets can be poisoned, leading to threats going undetected or false positive notifications.
- **Chatbot manipulation**—AI tools such as chatbots can be fed with poisoned datasets to produce inaccurate, hostile or offensive responses.
- **Health and safety exploitation**—Data poisoning attacks aimed at creating errors in autonomous driving systems or AI medical diagnostic tools can lead to significant injuries or fatalities.

Prevention Methods for Businesses

Given the far-reaching impacts of data poisoning attacks, businesses should consider these strategies to mitigate their exposure to them:

- **Data validation and sanitization**—Businesses should filter out potential attacks by removing data anomalies and investigating suspicious patterns. They should also verify the validity of the data sources used for training.
- **Secure data handling**—Utilizing encryption, access controls and secure protocols can add protective layers for datasets.
- **Monitoring and auditing**—Businesses should implement systems to detect vulnerabilities and data irregularities. This enables organizations to identify potential troublesome areas and address them before they create larger issues.
- **Diversification of data sources**—Training models on varied data sources can reduce attack risks by preventing targeted data manipulation.
- **Robust training techniques**—By incorporating adversarial training into model learning, AI can be equipped to handle tampered data.
- **Data provenance**—Maintaining clear records of data sources is beneficial when tracing to find potential points of compromise.
- **Output verification**—To quickly detect issues, businesses should frequently compare model outputs against expected behavior.
- **Comprehensive user training**—Raising awareness through ongoing training and education can help users recognize suspicious activity or outputs related to data poisoning.
- **Penetrative testing**—It's important to test systems regularly to see where they are vulnerable. This can help businesses proactively address weak points.

Conclusion

Data poisoning attacks pose serious risks. Businesses can reduce their exposure to these cybersecurity incidents by taking the time and initiative to implement prevention methods.

Contact us today for more information.