DATA POISONING: THE EXPLOITATION OF GENERATIVE AI

Bart Lenaerts-Bergmans - March 19, 2024

What is data poisoning?

Data poisoning is a type of cyberattack in which an adversary intentionally compromises a training dataset used by an AI or machine learning (ML) model to influence or manipulate the operation of that model.

Data poisoning can be done in several ways:

- Intentionally injecting false or misleading information within the training dataset
- Modifying the existing dataset
- Deleting a portion of the dataset

By manipulating the dataset during the training phase, the adversary can introduce biases, create erroneous outputs, introduce vulnerabilities (i.e., backdoors), or otherwise influence the decision-making or predictive capabilities of the model.

Data poisoning falls into a category of cyberattacks known as adversarial AI. Adversarial AI or adversarial ML is any activity that seeks to inhibit the performance of AI/ML systems by manipulating or misleading them.

Data poisoning symptoms

Because most AI models are constantly evolving, it can be difficult to detect when the dataset has been compromised. Adversaries often make subtle — but potent — changes to the data that can go undetected. This is especially true if the adversary is an insider and therefore has in-depth information about the organization's security measures and tools as well as their processes.

To spot a potential case of data poisoning, it is perhaps best to remember the reasons why most cybercriminals use this tactic: to decrease the accuracy, precision, and performance of the model. With that in mind, it's important to keep a look out for these warning signs of data poisoning:

Symptoms	Questions to ask
Model degradation	Has the performance of the model inexplicably worsened over time?
Unintended outputs	Does the model behave unexpectedly and produce unintended results that cannot be explained by the training team?
Increase in false positives/negatives	Has the accuracy of the model inexplicably changed over time? Has the user community noticed a sudden spike in problematic or incorrect decisions?
Biased results	Does the model return results that skew toward a certain direction or demographic (indicating the possibility of bias introduction)?
Breaches or other security events	Has the organization experienced an attack or security event that could indicate they are an active target and/or that could have created a pathway for adversaries to access and manipulate training data?
Unusual employee activity	Does an employee show an unusual interest in understanding the intricacies of the training data and/or the security measures employed to protect it?

Types of data poisoning

Data poisoning attacks are typically classified based on the intended outcome of the attack. The two most common categories of data poisoning are:

1. **Targeted data poisoning attacks**: Targeted attacks occur when an adversary is attempting to manipulate the model's behavior with respect to a specific situation. For example, a cybercriminal may train a cybersecurity tool to misidentify a specific file that they will use in a future attack or ignore suspicious activity from a certain user. Though targeted attacks can lead to serious and far-reaching consequences, they do not degrade the overall performance of an AI model.

2. **Non-targeted data poisoning attacks**: A non-targeted attack is when a cybercriminal manipulates the dataset to negatively impact the overall performance of the model. For example, the adversary may introduce false data, which in turn could reduce the accuracy of the model and negatively impact its predictive or decision-making capabilities.

EXPERT TIP

Internal vs. external actor Another key consideration when it comes to detecting and preventing data poisoning attacks is who the attacker is in relation to the target. In many cases, a data poisoning attack is carried out by an internal actor, or someone who has knowledge of the model and often the organization's cybersecurity processes and protocols. This is known as an insider threat, or white box attack.A black box attack, on the other hand, is carried out by an adversary that does not have inside information about the model they are attacking.Generally speaking, white box attacks tend to have a higher probability of success and cause more significant damage, underscoring the importance of protecting the organization from insider threats.

Examples of data poisoning attacks

With the broad categories of data poisoning attacks established, let's take a look at some specific tactics and techniques used by cybercriminals:

Backdoor poisoning

Backdoor poisoning involves injecting data into the training set with the intention of introducing a vulnerability that will serve as an access point, or "backdoor," for an attacker. The attacker can then use this point can to manipulate the model's performance and output. Backdoor poisoning can be either a targeted or non-targeted attack, depending on the specific goals of the attacker.

Availability attack

An availability attack is a type of cyberattack that attempts to disrupt the availability of a system or service by contaminating its data. Adversaries may use data poisoning to manipulate the data in a way that would degrade the performance or functionality of the targeted system, such as by making the system produce false positives/negatives, fail to process requests efficiently, or even completely crash. This would render the application or system unavailable or unreliable for its intended users.

Model inversion attacks

A model inversion attack uses the model's responses (its output) to recreate the dataset or generate assumptions about it (its input). In this type of attack, the adversary is most commonly an employee or other approved system user since they need access to the model's outputs.

Stealth attacks

A stealth attack is an especially subtle form of data poisoning wherein an adversary slowly edits the dataset or injects compromising information to avoid detection. Over time, the cumulative effect of this activity can lead to biases within the model that impact its overall accuracy. Because stealth attacks operate "under the radar," it can be difficult to trace the issue back through the training dataset, even after the issue is discovered.

The impact on AI

As organizations develop and implement new traditional and generative AI tools, it is important to keep in mind that these models provide a new and potentially valuable attack surface for threat actors. In the rush to capitalize on these new tools or test their usefulness, many teams may inadvertently overlook or underestimate the security of their models. Keeping security in mind is critical, even when using private large language models (LLMs) that are exclusive to the organization.

It is also important to remember that an adversarial AI attack — and data poisoning in particular — can have long-lasting and far-reaching implications. This is because the training data used by the model is compromised, which means that the output of the model can no longer be trusted.

If and when a breach is detected, organizations must attempt to trace the corruption back and restore the dataset. This requires a detailed analysis of the model's training data as well as the ability to scrub any false inputs and restore deletions. This is often impossible to do — but even in cases where it is possible, it is very time-

consuming and costly. In some cases, the model may have to be retrained completely, which is generally even more time- and resource-intensive.

Data poisoning of AI models can have potentially devastating consequences if a critical system is compromised and the attack goes undetected. For example, autonomous vehicles are controlled by AI systems; if the underlying training data is compromised, the decision-making capabilities of the vehicle could be impacted, potentially leading to accidents. Similarly, the use of AI in healthcare, financial services, and even utility systems opens the world up to significant risk.

Data poisoning defense best practices

Some data poisoning best practices include:

Data validation

Since it is extremely difficult for organizations to clean up and restore a compromised dataset after a data poisoning attack, prevention is the most viable defensive strategy. Organizations should leverage advanced data validation and sanitization techniques to help detect and remove anomalous or suspicious data points before they are incorporated into the training set.

Monitoring, detection, and auditing

AI/ML systems require continuous monitoring to swiftly detect and respond to potential risks. Companies should leverage cybersecurity platforms with continuous monitoring, intrusion detection, and endpoint protection. Models should also be regularly audited to help identify early signs of performance degradation or unintended outcomes.

Additionally, you have the option to incorporate live monitoring of input and output data into your AI/ML infrastructure. This involves scrutinizing the data continuously to detect any anomalies or deviations. By promptly identifying such irregularities, you can swiftly implement security measures to safeguard and fortify your systems against potential threats.

Continuous monitoring can also lead to the application of user and entity behavior analytics (UEBA), which you can use to establish a behavioral baseline for your ML model. Based on this, you can more easily detect anomalous patterns of behavior within your models.

Adversarial training

Adversarial training is a defensive algorithm that some organizations adopt to proactively safeguard their models. It involves introducing adversarial examples into a model's training data to teach the model to correctly classify these inputs as intentionally misleading.

By teaching an ML model to recognize attempts to manipulate its training data, you train the model to see itself as a target and defend against attacks such as model poisoning.

Data provenance

Organizations should retain a detailed record of all data sources, updates, modifications, and access requests. Though these features won't necessarily help detect a data poisoning attack, they are invaluable in helping the organization recover from a security event and identify the individuals responsible.

In the case of white box attacks, simply having robust data provenance measures in place can be a valuable deterrent.

Secure data handling

Establish and enforce clear and robust access controls for who has access to data, especially sensitive data. Employ the principle of least privilege (POLP), which is a computer security concept and practice that gives users limited access rights based on the tasks necessary for their job. The POLP ensures only authorized users whose identities have been verified have the necessary permissions to execute jobs within certain systems, applications, data, and other assets.

Organizations should also employ comprehensive data security measures, including data encryption, data obfuscation, and secure data storage.

User awareness and education

Many of your staff members and stakeholders may be unaware of the concept of data poisoning, let alone its threats and signs. As a part of your overall cybersecurity defense strategy, raise awareness through training programs and education. Train your teams on how to recognize suspicious activity or outputs related to AI/ML-based systems. You should also ask your security vendor how they harden their technology against adversarial AI. One way CrowdStrike fortifies ML efficacy against these types of attacks is by red teaming our own ML classifiers with automated tools that generate new adversarial samples based on a series of generators with configurable attacks.

When your staff is equipped with this kind of knowledge, you add an extra layer of security and foster a culture of vigilance that enhances your cybersecurity efforts.

CrowdStrike's approach

CrowdStrike is uniquely positioned to lead the security industry as it adopts generative AI. The AI-native CrowdStrike Falcon® platform has been at the forefront of AI-powered detection innovation since its founding.

To enable organizations to safely embrace generative AI, we've centered the needs and concerns of security teams in the architecture of **CrowdStrike® Charlotte AI™**, CrowdStrike's generative AI security analyst.

Three key features that set Charlotte AI apart:

- **Trusted data**: Charlotte AI uses high-fidelity intelligence contained in the Falcon platform, which provides built-in safeguards against data poisoning.
- **Auditable, traceable answers**: Every answer Charlotte AI provides can be inspected and audited using the "Show Response Details" toggle.
- **User education**: Charlotte AI supports ongoing upskilling of security team members and improves the employee experience by automating routine, recurring tasks.

Bart is Senior Product Marketing Manager of Threat Intelligence at CrowdStrike and holds +20 years of experience in threat monitoring, detection and intelligence. After starting his career as a network security operations analyst at a Belgian financial organization, Bart moved to the US East Coast to join multiple cybersecurity companies including 3Com/Tippingpoint, RSA Security, Symantec, McAfee, Venafi and FireEye-Mandiant, holding both product management, as well as product marketing roles.

Featured Articles

Data Gravity

Types of Cyberattacks

Data Privacy

English (US)