FEATURE FOCUS

AI DATA POISONING

By: Celine Gravelines

Artificial Intelligence (AI) models rely heavily on the quality and accuracy of their training data which, if subject to attacks such as data poisoning, compromises the effectiveness and reliability of the model itself, leading to potentially severe consequences. Security practitioners must prepare for these threats as AI is increasingly incorporated into standard business practices. This article discusses methods of data poisoning and broader implications of such attacks to the economy, health and safety, and public trust, while presenting a layered strategy to mitigate the risks.

AI Data Poisoning: Attack Methods and Mitigation Strategies

With the rise of AI, the evolving threat landscape requires security professionals to proactively understand the latest attack methods and how best to mitigate the risk and minimize impact, while enabling new innovative business functions. A study from Microsoft identified poisoning attacks as the highest ranked machine learning (ML)-related vulnerability perceived by 28 organizations in the industry [1], indicating data poisoning is a large deterrent for deploying these models. AI depends heavily on the quality of its training data, making that data an attractive asset to manipulate through attacks such as data poisoning. Data poisoning is an adversarial attack in which the training data used to train an ML model is compromised [2]. By manipulating the training data, adversaries can degrade the overall performance of the model (affecting its effectiveness and availability) or are able to create vulnerabilities which can be exploited when used in production (affecting the integrity). The training phase can be enormously expensive to conduct in the first place. OpenAI reported that the training of GPT-4 cost over $100 million USD [3], demonstrating that compromise of training data would be a major financial setback. Data poisoning attacks happen during the training phase of the machine learning algorithm while similar attacks occurring during the interference phase (i.e., once the model is deployed in production) are classified as evasion attacks [2].

Training data can be compromised in a variety of ways, including via insider threat (e.g., a rogue employee), a compromised system (e.g., external exploitation of infrastructure vulnerabilities to access training database), or a supply chain attack (e.g., compromise of third-party provider that supplies training data). Once access to the training data is available, adversaries can conduct various data poisoning methods to compromise the data, such as the following:

- **Label flipping attacks:** tampering with the ground truth labels associated with the training set, causing the model to be trained on inaccurate data, ultimately compromising the model's integrity and effectiveness (e.g., mislabeling spam email as innocuous to bypass spam filters) [4].

- **Clean-label attacks:** introducing subtle variances to the training data, undetectable to humans, without changing the labels, resulting in incorrect associations [5].

- **Data injection attacks:** introducing new, malicious data to the training data to mislead it's the learning. These attacks can be direct (to alter the behavior of the model) or indirect (to affect another system operation, similar to a SQL injection) [2].

- **Backdoor attacks:** injecting a specific pattern or "trigger" to the training data, so the model learns to associate this trigger with a particular label, regardless of the rest of the data's content, effectively creating a backdoor [2]. This attack is commonly used in computer vision applications where an image can include an inconspicuous trigger and regardless of the rest of the image itself, the model produces the malicious output if the trigger is present.

- **Availability attacks:** inserting random noise or outliers to the training data with the goal of reducing the overall performance, rendering the model unreliable and unusable (essentially causing a denial-of-services attack) [2].

Fraud Scenario

Consider an AI-based fraud detection platform used by a financial firm. Trained on vast amounts of transaction data, the model is intended to identify and differentiate between fraudulent and legitimate transactions. In this case, an adversary wants to commit fraud without being detected. To conduct a data poisoning attack, the adversary has gained access to the database containing the training data and plans to tamper with it to force the model to learn incorrect patterns. With a baseline of legitimate activities already existing in the training data (reflecting regular transactions, such as paying for groceries or money transfers to friends), he can begin to introduce minor anomalies that appear otherwise normal, just with slight deviates from that established regular behavior, such as paying new bills with slightly higher amounts.

By now, the system is accustomed to these minor anomalies during its training and the adversary can mimic performing more complex small-scale fraudulent transaction, taking advantage of normal spending habits, still designed to blend into the existing data patterns. These false negatives may not be flagged as illegitimate as they fall within the expanded range of normal behavior. As these models are always learning from new data, the model is now being trained to accept these transactions as legitimate. Over time, with more data poisoning, this range can expand even further. Once the model is put into production, it will not recognize certain transactions as fraudulent, which adversaries can take advantage of to have their fraudulent transactions go undetected.

Broad Implications

As demonstrated, data poisoning attacks can have broad economic impacts, resulting in major financial losses directly from fraud. Considering other applications, health and safety risks are a concern if, for example, autonomous cars are manipulated to ignore stop signs [6] or AI-based health imaging software learns to ignore tumors or suggest ineffective treatments. Likewise, data poisoning tactics can be used to manipulate viewing recommendations of streaming services to spread misinformation. More broadly, as attacks of this nature impact the effectiveness of AI systems, public confidence and trust in AI can be hampered, potentially slowing down adoption and innovation.

Mitigation Strategies

Unlike many other traditional vulnerabilities, these attacks are not exploiting a simple “bug” in the system that can be resolved with a patch or code fix. These adversarial attacks are subtle and difficult to detect by both humans and machines. A multi-layered defense, with security-by-design proactively implemented, provides the best strategy to mitigate the associated risks. Consider the following recommendations:

- **Data validation:** Verify the integrity and accuracy of the training data by performing consistency checks, completeness checks, confirm correctness against known standards, and ensure all values fall with acceptable ranges, limits, and formats. Consider using trusted data validation libraries and tools to automate the process.

- **Data sanitization:** Employ policy and protocols to regularly remove or correct inaccuracies, inconsistencies, duplicates, and potential malicious data. Proactively implement quality filtering to filter out low-quality or irrelevant data.

- **Anomaly detection:** Implement statistical models or pattern recognition systems for anomaly detection, to identify and flag unusual patterns and outliers in the training data.

- **Diverse data sources:** Relying on one source of training data can be a single point of failure if that source is compromised. Rather, leverage multiple datasets from different sources of creditable providers.

- **Data augmentation:** Introduce benign variations of the training data, such as rotations, translations, or cropping of the original datasets to mitigate the impact of tampered data without affecting the model’s performance [7]. By learning from a broad range of variations, the model can better generalize and be less susceptible to data poisoning attacks.

- **Access control:** Restrict access of training data to only authorized personnel. Employ policies and controls, such as multi-factor authentication and privileged access management, to reduce the risk of data tampering (whether intentional or unintentional).

- **Incident response:** Formalize incident response procedures to prepare for potential attacks like data poisoning. Consider steps to identify, contain, eradicate, and remediate efficiently, while also considering communication protocols with stakeholders.

- **Red-teaming:** Organize a team with varying skillsets to attempt to compromise the training data in a controlled environment. By simulating adversarial attacks, systematically discover vulnerabilities and potential threats, and test resilience. Track all findings for further testing, risk analysis, and remediation.

- **Monitoring and auditing:** Continuously monitor the models to detect unusual behavior or performance issues. Track metrics, set alerts, and leverage automated tools to identify potential vulnerabilities and compromises. Maintain records of data sources, data access, and modifications.

- **Third-party security:** If using training data from a third-party, verify the quality and accuracy of that data and consider what security controls the supplier has in place. Always leverage trusted, vetted sources for both the training data and model itself. Alternatively, if simply using third-party AI systems in business processes, validate the protection measures implemented by the vendor to mitigate against data poisoning and other risks.

- **Information sharing:** Collaborate with other organizations, government bodies, and research institutions facing similar AI risks to stay informed of emerging threats, trends, and lessons learned.

- **Other best practices:** Adhere to industry standards and best practices for both traditional security and AI security. For example, leverage the NIST AI Risk Management Framework [8] to formalize an approach to govern, map, measure, and manage AI-associated risks.

As many businesses are shifting to incorporate AI into aspect of their operations, services, and product offerings, security practitioners and developers alike have a responsibility to be aware of the associated risks and mitigation recommendations. Data poisoning can be insidious and nearly impossible to detect without strict measures in place. While preventative controls are the best first line of defense, ensure a holistic strategy to respond and remediate as AI threats continue to evolve.

About the Author

Céline Gravelines is a cybersecurity professional with 10 years of experience, specializing in data protection, incident response, policy, risk, privacy, and application security of both traditional and AI systems. She holds an MSc. in Computer Science, where she developed a novel application of unsupervised machine learning to map functional neurological data.

References

[1] R. S. Siva Kumar *et al.*, "Adversarial Machine Learning-Industry Perspectives," *2020 IEEE Security and Privacy Workshops (SPW)*, 2020. https://ieeexplore.ieee.org/document/9283867

[2] National Institute of Standards and Technology. “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.” NIST AI 100-2e2023, 2023. https://doi.org/10.6028/NIST.AI.100-2e2023

[3] Metz, Cade. "OpenAI CEO Sam Altman: The Age of Giant AI Models Is Already Over." *Wired*, May 16, 2023. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[4] E. Rosenfeld *et al.,* “Certified robustness to label-flipping attacks via randomized smoothing,” *Proceedings of the 37th International Conference on Machine Learning*, 2020. https://dl.acm.org/doi/abs/10.5555/3524938.3525700

[5] A. Gupta *et al.,* “Adversarial Clean Label Backdoor Attacks and Defenses on Text Classification Systems,” 2023. https://doi.org/10.48550/arXiv.2305.19607

[6] C. Sitawarin *et al.,* “DARTS: Deceiving Autonomous Cars with Toxic Signs.” 2018. https://arxiv.org/pdf/1802.06430v1

[7] E. Borgnia *et al.*, "Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks Without an Accuracy Tradeoff," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. https://ieeexplore.ieee.org/abstract/document/9414862

[8] National Institute of Standards and Technology. “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligent Profile.” NIST AI 600-1 Initial Public Draft, 2024. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

INFORMATION SYSTEMS SECURITY ASSOCIATION
40
40 YEARS STRONG
Assessing Your Cloud Service Providers: Impact of Security and Privacy in the Cloud
PRIVACY SPECIAL INTEREST GROUP WEBINAR SERIES
Live Webinar
August 20, 2024
1 PM ET
REGISTER NOW