Quinte
Financial Technologies®

THE EMERGING THREAT OF **AI POISONING**

Vectors, Risks, and Mitigation Strategies for Financial Institutions

19 Fulton Street, New York, New York 10038
(646) 813-0694
www.quinteft.com

INDEX

Introduction	03-04
AI Poisoning Mitigation Strategies	05-06
Assessing AI Poisoning Risk	07
Integrating AI Poisoning Detection in Practice	08
DATA ANALYTICS ServiceDESK The Analytics Edge for Trustworthy AI	09
Conclusion	10

INTRODUCTION

AI poisoning is an intentional attack aimed at corrupting the training data or models of artificial intelligence systems to undermine their integrity and performance. It involves manipulating training datasets or inserting harmful updates during the model training process, causing the AI to learn incorrect patterns, make consistent errors, or exhibit hidden malicious behaviors. The effects of AI poisoning can be severe, especially in critical sectors like healthcare, finance, and security, where a poisoned AI system could misdiagnose, make faulty financial predictions, or allow security breaches.

AI poisoning attacks fall into three primary categories:

- **Data Poisoning:** Malicious data is inserted into the training dataset to skew the model's learning, leading to errors or biases. Example: altered images in an autonomous vehicle's training data causing misclassification of traffic signs.

  In the financial domain, credit scoring models or fraud detection systems can be targeted when an insider or attacker injects false transaction patterns labeled as "normal," biasing the AI into misclassifying fraudulent activities as legitimate. One case involved a trader consistently tagged as "no fraud" who later engaged in fraud. The AI continued predicting "no fraud" because it had overfitted to the trader's past behavior, illustrating how poisoning can help fraudsters bypass detection.

- **Model Poisoning:** Harmful updates are injected during collaborative training, degrading the model or embedding vulnerabilities. In federated learning environments used by financial institutions, attackers can inject malicious updates to reduce model accuracy or create hidden weaknesses. This can lead to misclassification in risk assessments or compliance monitoring, creating systemic financial risks or potential compliance failures.

- **Backdoor Attacks:** Hidden triggers are embedded so the model behaves incorrectly when the trigger is present, such as granting unauthorized access when a specific pattern appears. In fraud detection, a backdoor could allow attackers to execute transactions with a special pattern (e.g., a specific sequence of amounts or timing) that the AI classifies as safe, enabling undetected fraud.

Attackers often exploit large datasets, open environments, and weak data integrity controls. Poisoned AI models can produce subtle but dangerous flaws that are hard to detect and may cause erroneous decisions or security vulnerabilities. This is increasingly concerning as AI systems become more widespread in high-stakes decision-making.

In financial services, consequences include inaccurate credit decisions, overlooked fraud or AML alerts, regulatory non-compliance, and reputational damage. Strong data governance, continuous model auditing, adversarial robustness training, and anomaly detection are essential to reducing these risks.

The examples above illustrate the practical and potentially damaging ways AI poisoning manifests in finance, underscoring the need for multidisciplinary mitigation strategies.

AI POISONING MITIGATION STRATEGIES

To mitigate AI poisoning, the National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF) emphasizes a comprehensive, proactive approach across data, models, and governance. Key mitigation strategies include:

Data Security and Validation

- Source training data only from trusted, verifiable repositories with strong provenance tracking.
- Implement automated data validation pipelines to sanitize, filter, and flag anomalies or duplicates.
- Continuously monitor datasets for suspicious changes or potential contamination.

Model Robustness and Monitoring

- Use adversarial and defensive training to enhance model resilience.
- Routinely audit and test model behavior for anomalies or poisoning indicators.
- Deploy anomaly detection systems that monitor both training and inference phases.

Access Controls and Governance

- Enforce strict role-based access control (RBAC) and least-privilege access.
- Maintain detailed audit logs for data usage and model updates.
- Control model fine-tuning and updates using pre-approved protocols and rollback options.

Incident Response and Collaboration

- Develop incident response plans and playbooks specifically for poisoning scenarios.
- Facilitate third-party audits to identify poisoned states early.
- Engage external stakeholders for threat-intelligence sharing and community feedback.

Emerging Technologies and Practices

- Use blockchain and federated learning to secure decentralized training and enhance transparency.
- Use synthetic or augmented data to reduce reliance on vulnerable real-world datasets.
- Apply AI-specific poisoning detection tools within CI/CD pipelines.

These layered defenses reflect the NIST AI RMF's principles of govern, map, measure, manage, and improve, aimed at building trustworthy AI systems.

ASSESSING AI POISONING RISK

According to the NIST AI RMF, assessing poisoning risk is part of the “Measure” function, which focuses on analyzing, assessing, benchmarking, and continuously monitoring AI risks.

NIST AI RMF outlines several key recommendations:

- Selecting quantitative, qualitative, or mixed assessment methods tailored to detecting data poisoning, model poisoning, and backdoor risks.
- Evaluating trustworthiness attributes of security, resilience, reliability, and fairness by testing for anomalies or behavior changes.
- Applying continuous tracking mechanisms during training and deployment to detect emerging poisoning attempts.
- Benchmarking AI performance against trusted baselines to catch deviations.
- Documenting assessment results, uncertainties, and detected risks.
- Using insights to prioritize mitigation actions and improve controls.

This assessment approach is iterative and should be embedded into a continuous risk management cycle that covers governance, context mapping, and proactive prevention.

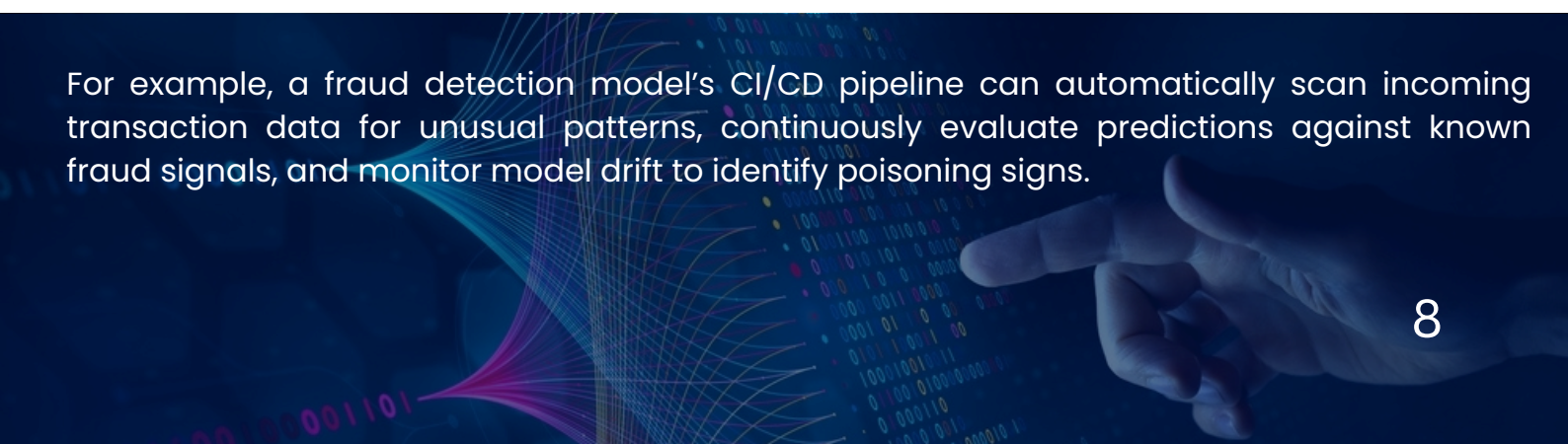
INTEGRATING AI POISONING DETECTION IN PRACTICE

Integrating poisoning detection into CI/CD pipelines requires embedding automated security checks throughout model development, training, and deployment to detect poisoning early.

Key methods include

- **Data Validation and Filtering:** Automatically vet all incoming data using anomaly detection and provenance verification.
- **Adversarial Training and Testing:** Incorporate adversarial examples into training and test models during builds for abnormal behavior.
- **Ensemble Modeling:** Use ensembles where independent models vote on predictions; disagreements may indicate poisoning.
- **Continuous Monitoring:** Monitor model behavior and data distributions in real time for anomalies or performance drops.
- **Canary Models and Shadow Testing:** Test lightweight canary models or shadow deploy new models alongside production to detect suspicious behavior.
- **Access Controls and Audit Trails:** Enforce tight controls and maintain detailed logs to trace poisoning sources.
- **Automated Rollbacks:** Revert to trusted model versions when poisoning is detected.

For example, a fraud detection model's CI/CD pipeline can automatically scan incoming transaction data for unusual patterns, continuously evaluate predictions against known fraud signals, and monitor model drift to identify poisoning signs.



DATA ANALYTICS ServiceDESK : THE ANALYTICS EDGE FOR TRUSTWORTHY AI

Our Data Analytics ServiceDESK helps institutions build the governance, oversight, and intelligence infrastructure needed to detect, contain, and recover from AI poisoning threats.

Working within your existing environment and without displacing current systems, ServiceDESK delivers outcomes that directly support poisoning resilience:

Outcomes

- **Cleaner data, reduced attack surface:** Improved data quality, lineage tracking, and validation controls reduce the exposure points through which poisoned data enters training pipelines.

- **Earlier threat detection:** Anomaly identification across transactions, models, and operational activity surfaces poisoning indicators before they compound into systemic risk.

- **Stronger compliance and audit readiness:** Consistent, documented data governance supports regulatory obligations and provides the audit trail needed to investigate and demonstrate model integrity.

- **More confident, defensible decisions:** Unified intelligence across fraud, credit risk, and operations ensures that AI-driven outputs remain reliable and explainable even under adversarial conditions.

- **Sustained AI performance at scale:** MLOps governance, monitoring, and retraining workflows keep models accurate and trustworthy as threats and data environments evolve.

By pairing sound AI security practices with a reliable analytics foundation, financial institutions can act faster, govern more effectively, and maintain the trust that AI-driven decisions demand.

CONCLUSION

AI poisoning is an evolving threat that compromises the integrity and reliability of AI systems in critical industries. By corrupting training data or models, attackers can introduce subtle vulnerabilities that lead to biased decisions, compliance failures, and systemic risks.

Mitigation requires a proactive, layered defense aligned with frameworks such as NIST AI RMF, including strong data governance, adversarial resilience, continuous monitoring, and strict access controls. Integrating automated detection within CI/CD pipelines ensures early identification and containment of poisoning.

As AI adoption expands, organizations must embed security, transparency, and resilience into AI systems. Collaboration across governance, technology, and community intelligence is essential to defending against poisoning risks and sustaining trust in AI-driven decisions.

Supporting this effort, Quinte's Data Analytics ServiceDESK helps financial institutions build the foundation that trustworthy AI requires: cleaner data, earlier threat detection, stronger compliance controls, and models that remain accurate and defensible as threats evolve. The result is an institution better equipped to govern its AI responsibly, respond to emerging risks with confidence, and maintain the integrity of every decision its models drive.

www.quinteft.com

Quinte Financial Technologies®

References:

1. Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
2. Weaponizing Data: AI Data Poisoning Threats - Akitra https://akitra.com/cybersecurity-implications-of-data-poisoning-in-ai-models/
3. What Is Data Poisoning? | CrowdStrike https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/data-poisoning/
4. What Is Data Poisoning? | IBM https://www.ibm.com/think/topics/data-poisoning
5. What Is Data Poisoning? [Examples & Prevention] - Palo Alto Networks https://www.paloaltonetworks.com/cyberpedia/what-is-data-poisoning
6. Protecting AI Systems from Data Poisoning Attacks - Sidechain Security https://sidechainsecurity.com/protecting-ai-systems-from-data-poisoning-attacks/
7. AI-Driven DevSecOps For Intelligent CI/CD Pipeline | Aviator https://www.aviator.co/blog/ai-driven-devsecops-building-intelligent-ci-cd-pipelines/
8. OWASP Top 10 for LLMs in 2025: Risks & Mitigations Strategies https://strobes.co/blog/owasp-top-10-risk-mitigations-for-llms-and-gen-ai-apps-2025/
9. LLM04:2025 Data and Model Poisoning - OWASP Gen AI Security Project https://genai.owasp.org/llmrisk/llm04-model-denial-of-service/
10. AI Security Challenges and Best Practices for 2025 https://www.invicti.com/blog/web-security/ai-security-challenges-best-practices-for-2025