AI is only as good as its data. Is your organization protecting its models?

AI Data Poisoning

What: AI data poisoning is a type of cyberattack where *adversaries intentionally introduce corrupt or deceptive data* into an AI system's training or operational data pipeline.

Why: To *deliberately influence key decisions* when humans use the AI model to support decision-making (e.g., national defense, healthcare, finance)

This attack manipulates the model's learning process, so that *the AI's outputs are aligned with the adversary's goals*.

Preventing AI Data Poisoning

Secure Data Pipelines

- Vet and verify all training data sources *before* ingestion.
- Implement *strict access controls* to prevent unauthorized modifications.
- Use *cryptographic techniques* such as hashing and digital signatures to ensure data integrity.
- Log and audit *data provenance* to track its origin and modifications.

Preventing AI Data Poisoning

Anomaly Detection & Continuous Monitoring

- Monitor *AI behavior* for unexpected shifts or performance degradation.
- Use automated anomaly detection tools to *flag inconsistencies* in model outputs.
- Track *data drift*, as unexpected changes in real-world data can indicate poisoning attempts.
- Set up *real-time alerts* for unusual patterns in AI decision-making.

Preventing AI Data Poisoning

Adversarial Testing & Model Robustness

- Simulate data poisoning attacks during AI development to *identify vulnerabilities*.
- Conduct *red-team testing* on AI models to find weaknesses before attackers do.
- Use *differential privacy techniques* to make AI models less sensitive to small data manipulations.
- Retrain models periodically using *verified datasets* to correct potential poisoning effects.