CYBEATS

Vulnerabilities and Weaknesses of AI

Navigating the Challenges in Artificial Intelligence

Dmitry Raidman

May 2024

Introduction to AI Vulnerabilities

What Are AI Vulnerabilities?

AI vulnerabilities are weaknesses in artificial intelligence systems that can lead to incorrect, unethical, or harmful outcomes due to flaws in data handling, model design, infrastructure, or ethical guidelines.

Types of AI Vulnerabilities

Common Vulnerabilities in AI Systems

- Poisoning the Training Data (Misclassification)
- Model Poisoning (Targeting availability and Integrity)
- Neural Trojan Attacks (Specific input Conditions that affect model output)
- Model Transfer Attacks (Supply chain)
- Malicious Model Files and Dependencies (Supply Chain)
- Dependency Confusion and Typosquatting (AI Hallucination)