AI Security Threats

https://linkmyte.ai/e2Ud3364Hdg

Masood I. Khan

A-CISO, CRISC,CISA, CCSK, CDPP, SCCP, PMP, PMI-RMP, PECB Certified Trainer, ISO/IEC 27001 Senior Lead Implementer, ISO/IEC 27001 Senior Lead Auditor, ISO 31000 Senior Lead Risk Manager, ISO 22301 Lead Implementer, ITIL(F), MCTS, C-KPIP, LSSBB

Threat	Category	Description	Examples	Impact	Mitigation
Data Poisoning	Training Phase	Injecting malicious or incorrect data into training sets to corrupt model behavior and outcomes	Label flipping, clean-label backdoors, biased data insertion	**Critical** Compromises model integrity from inception	Data sanitization, anomaly detection, robust training algorithms, provenance tracking
Model Inversion	Privacy Attack	Extracting sensitive information used for training by querying the model	Reconstructing user data from model outputs, recovering facial images from recognition models	High Privacy violations and data breaches	Differential privacy, output perturbation, access controls, query monitoring
Adversarial Attacks	Input Manipulation	Creating special inputs (e.g., images or text) to mislead AI model predictions	Slightly modified stop sign fools self-driving car, perturbed audio in speech recognition	High Can cause critical misclassifications in autonomous systems	Adversarial training, input validation, ensemble methods, certified defenses
Model Stealing/Extraction	Intellectual Property	Duplicating a proprietary model by querying it repeatedly to extract intellectual property	API query attacks, side-channel attacks, model reverse engineering	High Loss of competitive advantage and IP theft	Query rate limiting, API monitoring, watermarking, differential privacy
Privacy Leakage	Privacy Attack	AI outputs reveal private or training data unintentionally	Memorized PII in language models, training data extraction from outputs	High GDPR violations, exposure of sensitive information	Data minimization, output filtering, differential privacy, regular audits
Backdoor Attacks	Training Phase	Embedding secret triggers that cause malicious behavior in models under specific conditions	Trojan triggers in models, recognizing illicit content after certain input	**Critical** Hidden vulnerabilities activated on demand	Model verification, neuron analysis, trigger detection, secure training pipelines
Evasion Attacks	Detection Bypass	Modifying inputs to evade AI-based security systems	Malware disguised to evade AI detection, spam filter bypass, deepfake detection evasion	High Renders security systems ineffective	Ensemble detection, adversarial training, multi-modal verification
Data Inference	Privacy Attack	Inferring confidential or protected data by analyzing AI output patterns	Deduction of user attributes from model responses, property inference	High Privacy concerns and compliance issues	Output noise injection, query restrictions, access controls
AI-Enhanced Social Engineering	Human Factor	Using GenAI to craft Highly convincing, targeted social engineering schemes	Deepfake phishing, spear-phishing, synthetic identity fraud, fake executive calls	High Increased success rate of fraud and manipulation	Multi-factor authentication, user awareness training, deepfake detection tools

Threat	Category	Description	Examples	Impact	Mitigation
API Attacks	Infrastructure	Exploiting AI model APIs for unauthorized access, data extraction, or denial of service	Input manipulation, API fuzzing, parameter tampering, injection attacks	High System compromise and data breaches	API security best practices, input validation, rate limiting, authentication
Hardware Vulnerabilities	Infrastructure	Exploiting weaknesses in AI processing hardware	Side-channel attacks, electromagnetic leaks, GPU memory access patterns	High Information leakage and system compromise	Secure enclaves, constant-time operations, hardware security modules
Model Poisoning	Supply Chain	Directly altering a deployed model, especially in federated learning or collaborative settings	Subtle distributed parameter changes, compromised model updates	**Critical** Widespread deployment of compromised models	Byzantine-robust aggregation, model verification, integrity checks, trusted sources
Membership Inference	Privacy Attack	Determining if specific data was used to train an AI model	Query-based membership inference, identifying if personal records were in training set	Medium Privacy concerns and regulatory compliance issues	Differential privacy, regularization techniques, confidence thresholding
Transfer Learning Attacks	Supply Chain	Adding malicious behavior or bias through manipulation of pre-trained models	Backdooring pre-trained vision models, poisoned foundation models	**Critical** Affects all downstream applications	Model provenance tracking, verification of pre-trained models, fine-tuning audits
Prompt Injection	Input Manipulation	Specially crafted prompts that manipulate LLMs or bypass safety controls	Jailbreak prompts, output manipulation, system prompt leaking, instruction override	High Can bypass safety measures and leak data	Input filtering, prompt isolation, output monitoring, instruction hierarchy
Supply Chain Attacks	Supply Chain	Compromising AI through tampered datasets, models, or third-party toolchains	Poisoned public datasets, malicious libraries, compromised model repositories	**Critical** Widespread systemic compromise	Supply chain verification, code signing, trusted repositories, dependency scanning
Automated Malware Generation	AI Misuse	Using AI to invent new malware or evade signature-based detection	GPT-written malware, AI-generated polymorphic code, automated exploit creation	High Accelerates malware development and evasion	Behavioral analysis, AI-powered defense systems, sandboxing
Deepfake Generation	AI Misuse	Producing convincing fabricated audio/video for fraud or manipulation	Fake executive voice calls, synthetic video for disinformation, identity fraud	High Fraud, reputation damage, political manipulation	Digital watermarking, deepfake detection tools, authentication protocols

Threat	Category	Description	Examples	Impact	Mitigation
Brute Force & DoS Amplification	Availability	AI-optimized attacks that speed up brute force or denial of service campaigns	Automated credential stuffing, botnet DoS, sponge examples, resource exhaustion	High System unavailability and resource exhaustion	Rate limiting, CAPTCHA, resource quotas, anomaly detection, load balancing
Inference Manipulation	Runtime Attack	Manipulating model inference process or deployment environment	Weight modification, runtime parameter tampering, memory corruption	**Critical** Direct control over model behavior	Secure enclaves, integrity monitoring, encrypted inference, runtime protection
Byzantine Attacks	Distributed Learning	Malicious participants in federated or distributed learning systems	Corrupted gradient updates, model poisoning in federated learning	High Compromises collaborative AI training	Byzantine-robust aggregation, participant verification, outlier detection