International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Artificial Intelligence for Cybersecurity: Threats, Defenses, and Emerging Challenges in the Era of Generative AI

Sonal Sinha
Digital India Corporation, MeitY, Electronics Niketan, CGO Complex, New Delhi, India

Abstract — Artificial Intelligence (AI) has a dual relationship with cybersecurity: it is both a powerful defensive technology and, increasingly, a high-value attack target in its own right. Machine learning, deep learning, foundation models, and large language models (LLMs) now underpin intrusion detection, malware analysis, phishing identification, and autonomous incident response. At the same time, adversaries exploit AI for automated reconnaissance, AI-generated phishing, and attacks aimed directly at AI models — adversarial evasion, data poisoning, model extraction, prompt injection, jailbreaking, and AI supply-chain compromise. This paper condenses a comprehensive survey into a conference-length treatment: it traces the evolution of AI in cybersecurity, presents a unified taxonomy of attacks across the AI lifecycle, reviews key defense mechanisms (explainable AI, federated learning, differential privacy, guardrails, zero trust), summarizes major governance frameworks (NIST AI RMF, ISO/IEC 42001, EU AI Act, OWASP LLM Top 10), and outlines open research challenges for building trustworthy, resilient AI-driven security systems.

Keywords— Artificial Intelligence; Cybersecurity; Generative AI; Large Language Models; Adversarial Machine Learning; Prompt Injection; Explainable AI; Federated Learning; AI Supply-Chain Security; Zero Trust; AI Governance.

I. INTRODUCTION

The convergence of cloud computing, mobile platforms, IoT, and cyber-physical systems has expanded the attack surface available to adversaries, making cybersecurity a matter of national security and economic stability. Traditional signature- and rule-based defenses struggle against advanced persistent threats, polymorphic malware, zero-days, and AI-generated phishing. Artificial Intelligence — spanning classical machine learning, deep learning, and now foundation models and LLMs — has become central to modern defense, powering intrusion detection, malware analysis, user/entity behavior analytics, and security orchestration platforms.

This relationship is inherently dual-use. The same generative capabilities that help defenders can produce convincing phishing content, deepfakes, and malicious code, and autonomous AI agents can be used to coordinate attacks. More concerning still, the AI systems themselves — their training data, model parameters, and deployment pipelines — are now attack surfaces: adversarial examples, poisoning, model extraction, and (for LLMs) prompt injection and RAG poisoning can silently corrupt or steal the very defense mechanism an organization relies on. Governance bodies have responded with frameworks such as the NIST AI Risk Management Framework, ISO/IEC 42001, the EU AI Act, and OWASP's LLM guidance, but much of the literature still treats AI-for-security and security-of-AI separately. This paper condenses a broader survey to unify both perspectives: Section II traces the evolution of AI in cybersecurity; Section III covers AI system vulnerabilities and threat models; Section IV presents an attack taxonomy; Section V reviews defenses; Section VI summarizes governance frameworks; Section VII outlines open challenges; Section VIII concludes.

II. EVOLUTION OF AI IN CYBERSECURITY (2015–2026)

AI's role in cybersecurity has progressed through five overlapping phases, summarized in Table I: an early machine-learning era built on handcrafted features and classical classifiers; a deep-learning phase that automated feature extraction but introduced adversarial vulnerability as a research concern; a trustworthy-AI phase emphasizing explainability (LIME, SHAP) and privacy-preserving training (federated learning, differential privacy); the generative-AI/foundation-model era, which brought powerful threat-intelligence and code-analysis capabilities alongside prompt injection and RAG poisoning; and, most recently, autonomous AI agents capable of planning and tool use, which introduce agent-specific risks such as memory poisoning and goal hijacking. Across all five phases, each gain in detection capability and automation has been matched by a new class of vulnerability, reinforcing that securing AI is now inseparable from using AI for security.

Table 1. Evolution of AI in Cybersecurity

Period	Dominant AI Technologies	Applications	Emerging Challenges
2015–2017	ML: SVM, Random Forest, Decision Trees, KNN	Malware detection, IDS, spam filtering, anomaly detection	Feature engineering, scalability, zero-day detection
2017–2020	DL: CNN, RNN, LSTM, Autoencoders, GANs	Behavioral analytics, malware classification, network IDS	Adversarial attacks, robustness, explainability
2020–2022	XAI, Federated Learning, Privacy-Preserving ML	Trustworthy AI, collaborative & secure analytics	Transparency, fairness, privacy, governance
2022–2024	Foundation Models, Transformers, LLMs	Threat intelligence, secure coding, phishing/malware analysis, SOC assist	Prompt injection, jailbreaks, RAG poisoning, misuse
2024–2026	Autonomous Agents, Multi-Agent Systems, AI-native platforms	Autonomous SOCs, AI-assisted pen-testing, adaptive defense	Agent security, supply-chain attacks, execution safety, compliance

III. VULNERABILITIES AND THREAT MODELS FOR AI SYSTEMS

Unlike conventional software, AI behavior is derived from training data, learned parameters, and continuous interaction with dynamic environments, so its attack surface spans the entire lifecycle rather than a fixed codebase. Data collection and preprocessing are exposed to poisoning, label manipulation, and privacy leakage; training is exposed to backdoor implantation and federated-learning poisoning; deployment is exposed to API abuse and model stealing; inference is exposed to adversarial examples, prompt injection, and jailbreaks; and continuous learning is exposed to online-learning poisoning and reward hacking.

Threat models for AI are typically classified by attacker knowledge: white-box (full knowledge of architecture, training data, and parameters — used mainly for robustness evaluation), black-box (only input-output access, the realistic case for most public APIs), and gray-box (partial knowledge, e.g., of the feature representation or model family, reflecting many real-world insider or intelligence-informed scenarios). Correspondingly, AI security objectives extend beyond classical confidentiality, integrity, and availability (CIA) to include privacy (resistance to membership inference, attribute inference, and reconstruction attacks) and trustworthiness (robustness, fairness, transparency, explainability, and accountability) properties increasingly mandated by regulatory frameworks rather than treated as optional design goals.

LLM-based systems add a further layer of risk because they interact via natural language and often integrate external tools, plugins, and retrieval pipelines: prompt injection, indirect prompt injection, jailbreaks, system-prompt leakage, RAG/vector-database poisoning, tool manipulation, and agent-memory poisoning all exploit model behavior rather than conventional software flaws, often bypassing standard security controls.

This is compounded by AI supply-chain risk — organizations increasingly depend on open-source pretrained models, public datasets, model repositories, and third-party plugins, any of which can propagate a compromise downstream.

IV. TAXONOMY OF ATTACKS ON AI AND LLM-BASED SYSTEMS

Attacks on AI systems can be organized into six categories by target and lifecycle stage: data-centric attacks (data/label/clean-label poisoning, backdoor data injection); model-centric attacks (adversarial examples, evasion, model stealing, model inversion, membership/attribute inference); LLM-specific attacks (prompt injection, jailbreaks, prompt leakage, hallucination exploitation, RAG poisoning, context manipulation); AI-agent attacks (tool manipulation, agent memory poisoning, goal hijacking, multi-agent exploitation); supply-chain attacks (poisoned pretrained models, malicious checkpoints, dependency confusion, repository compromise); and infrastructure attacks (API abuse, resource exhaustion, cloud/edge AI attacks, denial-of-service). Table II summarizes representative attacks, their target, lifecycle stage, primary impact, and typical defense.

Table 2. Representative Attacks on AI Systems

Attack	Target / Stage	Primary Impact	Typical Defense
Data Poisoning	Training data / Training	Corrupted learning	Data validation, provenance, robust training
Label Poisoning	Labels / Training	Misclassification	Dataset auditing
Adversarial Attack	Input samples / Inference	Prediction errors	Adversarial training, input sanitization
Model Extraction	Model / Deployment	IP theft	Query rate limiting, watermarking
Membership Inference	Training data / Deployment	Privacy leakage	Differential privacy
Prompt Injection	LLM / Inference	Policy bypass	Prompt isolation, input filtering
Jailbreak	LLM / Inference	Safety bypass	Safety tuning, output moderation
RAG Poisoning	Knowledge base / Retrieval	Incorrect responses	Trusted indexing, retrieval validation
Agent Memory Poisoning	AI agent / Continuous learning	Persistent compromise	Memory validation, integrity checks
Supply-Chain Attack	AI ecosystem / Development	System-wide compromise	Model signing, SBOMs, provenance verification

V. DEFENSE MECHANISMS AND TRUSTWORTHY AI

Defense spans the same lifecycle as the attacks it counters: secure data collection and provenance tracking; secure model development (adversarial training, defensive distillation, ensemble learning); secure deployment (access control, rate limiting, container hardening); and runtime monitoring for drift and anomalous query patterns. Privacy-preserving techniques — federated learning, differential privacy, secure multi-party computation, and homomorphic encryption — allow collaborative training without exposing raw data, at the cost of utility or computational overhead. Explainable AI (LIME, SHAP, attention-based attribution) supports analyst trust and regulatory transparency requirements. For LLMs specifically, prompt-isolation and input-filtering defenses, retrieval validation for RAG pipelines, and AI guardrails that enforce output-safety policies are the current state of practice, complemented by Zero Trust architectures that apply continuous verification and least-privilege access to AI services themselves. Table III compares the primary techniques.

Table 3. Defense Techniques for AI Systems

Technique	Threat Addressed	Strengths	Limitations
Adversarial Training	Evasion attacks	Improves robustness	Computationally expensive
Defensive Distillation	Adversarial examples	Smooths decision boundaries	Limited vs. adaptive attacks
Ensemble Learning	Model manipulation	Higher resilience/stability	Increased resource use
Federated Learning	Data privacy	Collaborative training, no raw-data sharing	Vulnerable to poisoned updates
Differential Privacy	Membership inference	Strong privacy guarantees	Reduced utility at high privacy levels
Homomorphic Encryption	Data confidentiality	Computation on encrypted data	High computational overhead
Explainable AI (XAI)	Lack of transparency	Trust and accountability	May add model complexity
AI Guardrails	Prompt injection / jailbreaks	Enforces output-safety policy	Requires continual tuning
Zero Trust Architecture	Unauthorized access	Continuous verification, least privilege	Complex to implement

VI. GOVERNANCE, STANDARDS, AND REGULATORY FRAMEWORKS

Governance frameworks translate the technical objectives above into organizational accountability. The NIST AI Risk Management Framework structures risk management around four functions — Govern, Map, Measure, and Manage — and is voluntary and technology-neutral. ISO/IEC 42001 defines a certifiable AI management-system standard. The EU AI Act imposes a comprehensive, risk-tiered legal regime within the EU. The OECD AI Principles and UNESCO's Recommendation on the Ethics of AI provide broad, non-

binding international consensus on responsible AI, the latter emphasizing human rights and societal impact. OWASP's Top 10 for LLM Applications gives practical, developer-facing guidance focused on application-level risks such as prompt injection and insecure output handling. Table IV summarizes these frameworks; in practice, organizations increasingly combine a risk-management framework (NIST/ISO) with domain-specific guidance (OWASP) and jurisdictional compliance (EU AI Act) rather than relying on any single standard.

Table 4. Major AI Governance Frameworks

Framework	Primary Focus	Key Strengths	Limitations
NIST AI RMF 1.0	Risk management	Flexible, lifecycle-based, tech-neutral	Voluntary
ISO/IEC 42001	AI management systems	International certification standard	Significant implementation effort
EU AI Act	Legal regulation	Comprehensive, risk-based	Primarily EU-applicable
OECD AI Principles	Responsible AI	Broad international consensus	Non-binding
UNESCO Recommendation	Ethics	Human-rights/societal perspective	Limited technical guidance
OWASP LLM Top 10	Application security	Practical developer guidance	Application-level scope only

VII. OPEN RESEARCH CHALLENGES AND FUTURE DIRECTIONS

\

Several challenges remain open across the areas surveyed above:

- Adversarial robustness: moving from empirical defenses toward certified robustness and formal verification against adaptive attackers.
- LLM and agent security: secure prompting, output guardrails, and bounded-autonomy/runtime governance for agents that plan and invoke tools.
- Explainability at scale: intrinsically interpretable architectures that do not trade away performance.
- Privacy-preserving learning: reducing the computational overhead of federated and confidential-computing approaches enough for production use.
- AI supply-chain assurance: AI-specific software bills of materials (AI-SBOMs), provenance verification, and trusted model repositories.
- Standardized benchmarking: common, reproducible measures of robustness and trustworthiness across models and vendors.
- Governance harmonization: reconciling voluntary frameworks (NIST, OECD) with binding regulation (EU AI Act) across jurisdictions.

Longer-term, the field is moving toward AI-native, largely autonomous security operations ("SOC 2.0"), post-quantum-safe AI architectures, and self-healing cyber-defense systems — all of which will need to be paired with secure-by-design engineering and continuous human oversight rather than treated as drop-in replacements for analyst judgment.

VIII. CONCLUSION

AI has transformed cyber defense while simultaneously becoming a first-class target: its training data, models, and deployment pipelines are now attack surfaces in their own right, and the rise of generative AI and autonomous agents has added prompt injection, RAG poisoning, and agent-memory attacks to the classical list of adversarial and poisoning threats. Building trustworthy AI-driven security requires a lifecycle-wide approach — robust training, privacy-preserving computation, explainability, supply-chain assurance, and continuous monitoring — integrated with governance frameworks such as the NIST AI RMF, ISO/IEC 42001, the EU AI Act, and OWASP's LLM guidance. Closing the remaining gaps in certified robustness, agent governance, and standardized benchmarking will determine how safely AI can be trusted to defend the digital infrastructure it increasingly puts at risk.

REFERENCES

1. NIST. AI Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023.
2. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. International Organization for Standardization, 2023.
3. European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024.
4. OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 2019 (updated 2024).
5. UNESCO. Recommendation on the Ethics of Artificial Intelligence, 2021.

6. OWASP Foundation. OWASP Top 10 for Large Language Model Applications, 2025.
7. Khaleel Khan Mohammed. "The Future is Cloud: Modernizing Big Data for the Cloud Era"