A SURVEY ON

Security threats to and from AI models

Authors:
Kateryna Mishchenko
Iraklis Symeonidis
Niclas Ericsson

RISE Report, December 20, 2025
Funded by:
Center for Cybersecurity at RISE
CitCom.ai - AI TEF SCC
and the STRIDE project

Abstract

Artificial intelligence (AI) systems introduce novel cybersecurity challenges both as vulnerable targets and as potential threat vectors. This report presents a state-of-the-art analysis of cybersecurity risks to AI, such as adversarial attacks, data poisoning, and model inversion, and risks from AI, including automated phishing, deepfake generation, and adaptive malware. It outlines the dual role of AI in cybersecurity, explores actor motivations ranging from nation-states to insiders, and reviews methodologies, taxonomies, and tools for AI-specific threat modeling and mitigation. The findings highlight the need for integrated approaches that combine robustness, transparency, and continuous monitoring to secure AI systems and prevent their misuse. The analysis is supported by a structured review of the literature of academic and gray sources, which reveals dominant themes such as generative AI, adversarial attacks, and governance gaps.

1 Introduction

Artificial intelligence (AI) is increasingly embedded in critical infrastructure and digital systems, offering transformative capabilities in all sectors. However, its integration introduces a distinct set of cybersecurity challenges that differ fundamentally from those found in traditional IT environments. Unlike rule-based systems, AI models are data-driven, adaptive, and often opaque, making them susceptible to emergent risks and novel attack vectors [1, 2, 3].

This report investigates cybersecurity threats both *to* and *from* AI systems. On one hand, AI models are vulnerable to adversarial manipulation, data poisoning, and model inversion [4, 5]. On the other hand, they can be weaponized to automate cyberattacks, generate misinformation, and amplify malicious capabilities [6, 7, 8].

Figure 1 shows the dual role of AI, as both a target and a threat vector, which requires a rethinking of cybersecurity strategies, combining conventional defenses with AI-specific robustness and governance mechanisms [9].

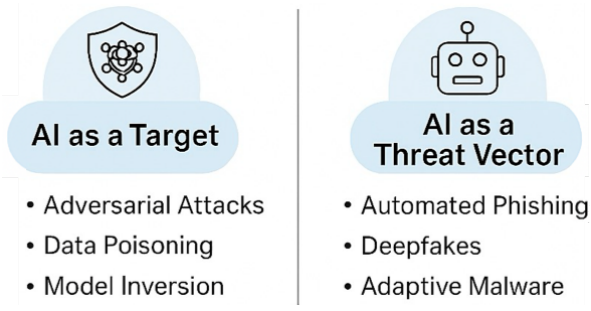

Figure 1: AI's dual role in cybersecurity, as a target and as a threat vector.

This report summarized findings from a structured review of academic and grey literature published between 2021 and 2025, covering both technical and governance aspects of AI security. It outlines the need for structured methodologies, taxonomies, and tools to assess and mitigate AI-related risks, and sets the stage for a multi-actor perspective, recognizing the diverse motivations behind AI exploitation, from nation-states and cybercriminals to insiders and design flaws. The following sections present the main categories of threats to and from AI, provide an overview of relevant frameworks and tools, and summarize the key insights that readers can apply in risk assessment, system design, and security planning.

2 Cybersecurity Threats to AI

AI systems differ fundamentally from traditional IT systems in how they process information, learn from data, and adapt to new contexts. This makes them vulnerable to a range of new and emerging cybersecurity threats that conventional defenses are not designed to handle [1, 2, 3].

One major category of threats involves **adversarial attacks**, in which inputs are intentionally crafted to fool the model into making incorrect decisions. These attacks can undermine the reliability of AI systems in critical applications such as autonomous control or anomaly detection [7, 10].

Another significant threat is **data poisoning**, which targets the training phase by injecting malicious data that corrupt the behavior of the model or embed hidden backdoors [2, 11]. These attacks are particularly dangerous because they can remain undetected until the model is deployed. This risk can be reduced by validating and filtering training data prior to ingestion, as demonstrated in studies showing how poisoned samples can alter model behavior and how proper data screening mitigates such manipulation [2].

Model inversion and **membership inference** attacks aim to reconstruct or identify sensitive data from model output, posing serious privacy risks when AI systems process personal or proprietary information [4].

Additionally, **data leakage** can occur when pri-

vate information is unintentionally exposed through model parameters or output. This is especially critical in regulated domains such as healthcare or finance [12, 4].

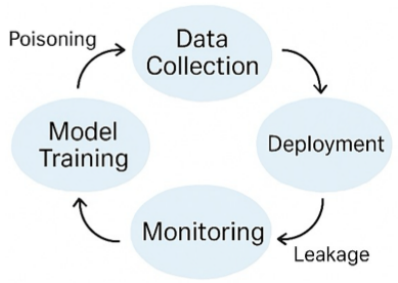

Figure 2: AI system lifecycle stages and vulnerabilities.

These threats compromise the confidentiality, integrity, and trustworthiness of AI systems. Mitigation strategies include adversarial training, robust data validation, explainability tools, and continuous monitoring [9, 5]. However, securing AI requires a shift in mindset, blending cybersecurity with machine learning robustness and responsible AI practices. For example, adversarial training can harden image-classification models by exposing them to perturbed inputs during training, a technique shown to significantly improve robustness against evasion attacks [3]. Other threats such as model inversion and membership inference require mitigation techniques that limit unintended information leakage. For instance, restricting confidence outputs and controlling model access can reduce exposure to model inversion attacks [12], while regularization and privacy-preserving mechanisms such as differential-privacy-inspired noise addition help decrease susceptibility to membership inference [4, 13]. As shown in Figure 2, vulnerabilities occur throughout the AI lifecycle, from data collection to deployment and monitoring.

3 Cybersecurity Threats from AI

While AI systems are vulnerable to attacks, they also pose significant risks as tools for cyber offense. Malicious actors increasingly exploit AI technologies to automate, scale, and enhance the effectiveness of cyberattacks [6, 14, 15].

One prominent threat is **automated phishing and social engineering**, where language models generate convincing and personalized messages at scale. For instance, generative AI systems have been shown to create highly personalized phishing emails that mimic internal communication patterns, significantly increasing click-through success rates. This reduces the barrier to launching widespread and targeted attacks, making traditional detection methods less effective [6, 16]. Recent incident reports have documented the use of large language models to craft targeted spear-phishing emails that successfully bypassed enterprise defenses, illustrating how AI-generated content is already being deployed in real-world attacks [15].

Deepfakes and synthetic media represent another growing concern. AI-generated videos, images, and voices can be used for impersonation, disinformation, and fraud. These techniques undermine public trust and can be weaponized to manipulate social discourse or defame individuals [15, 17].

Adaptive malware powered by AI can learn and evolve to evade traditional security measures. A concrete example is AI-generated malware that automatically mutates its code to bypass signature-based detection tools, a trend documented in recent analyses of emerging AI-enabled malware families [18]. Unlike static threats, these systems adapt over time, rendering signature-based detection tools obsolete [9, 18].

AI also accelerates **credential stuffing and vulnerability discovery**, enabling attackers to automate reconnaissance and exploit systemic weaknesses across platforms [5, 19].

In strategic contexts, AI is deployed in **autonomous weapon systems**, **mass surveillance**, and **cyberwarfare operations**. These applications raise ethical and regulatory concerns, especially when AI systems operate with limited human oversight [6, 20, 21].

The use of AI as a threat vector increases the scalability, stealth, and impact of cyberattacks. Defenders must adopt more sophisticated countermeasures, including adversarial testing, explainability tools, and continuous monitoring to mitigate these risks [9, 5].

4 AI as a Strategic Weapon

Beyond conventional cybersecurity threats, AI technologies are increasingly deployed in strategic and geopolitical contexts. These applications raise ethical, regulatory, and operational concerns, particularly when AI systems operate with limited human oversight [6, 21].

Autonomous weapon systems represent a major shift in military capabilities. AI enables drones and platforms to operate independently, transitioning from human-in-the-loop to human-on-the-loop configurations. For example, human-on-the-loop drone systems can autonomously select and track

targets, raising escalation risks if misclassification or communication failures occur during high-tempo operations [21]. This shift introduces accountability challenges and increases the risk of unintended escalation in conflict scenarios [6, 20].

Mass surveillance is another strategic use of AI, where facial recognition and behavioral monitoring systems are deployed by state actors. A well-documented example is the deployment of large-scale facial-recognition networks in authoritarian regimes to identify dissidents and monitor public behavior in real time, enabling systemic repression through automated surveillance [22]. These technologies can erode privacy, suppress dissent, and enable authoritarian control, especially in environments lacking regulatory safeguards [22, 23, 17].

AI also plays a role in **cyberwarfare operations**, where nation-states integrate AI into offensive and defensive cyber capabilities. This includes automated threat detection, strategic targeting, and adaptive response mechanisms. The integration of AI into national defense infrastructures introduces a new arms race dynamic in cyberspace, with unclear international norms and limited transparency [24].

These developments blur the line between civilian and military uses of AI, underscoring the need for robust governance frameworks, international cooperation, and ethical oversight.

5 Threat Actors and Motivations

Securing AI systems requires understanding the diverse actors who exploit or target them, each driven by distinct motivations and capabilities [9, 6, 25].

Nation-states pursue AI-driven capabilities for strategic, surveillance, and offensive purposes. AI enhances national defense mechanisms, enabling autonomous cyber warfare, automated threat detection, and mass surveillance. For example, national cyber units increasingly use AI-driven automation to accelerate vulnerability discovery and route reconnaissance, shortening the preparation time for offensive cyber operations [6]. These actors may exploit vulnerabilities in AI systems to gain geopolitical advantages or disrupt adversary infrastructures [24].

Cybercriminals use AI technologies to automate and scale traditional criminal activities such as ransomware, data theft, and fraud. AI-generated phishing emails, credential stuffing, and deepfake content increase the effectiveness and reach of these attacks [6, 19, 15].

Insiders, including employees or contractors with privileged access, may intentionally or unintentionally exploit AI vulnerabilities. Their access allows them to bypass external defenses and trigger attacks that leverage AI to evade detection [26, 27]. A simple example is an insider modifying training data or model parameters to degrade a system's accuracy or embed hidden behaviors, a manipulation that can be difficult to detect without robust auditing mechanisms [28].

Unintentional triggers arise from flawed models, insufficient training data, or lack of oversight. These can lead to unpredictable behavior or vulnerabilities that are later exploited by malicious actors. Such risks often stem from design flaws or inadequate testing rather than deliberate intent [1, 29].

Figure 3 shows the main categories of threat actors, nation-states, cybercriminals, insiders, and unintentional triggers, along with their primary motivations.

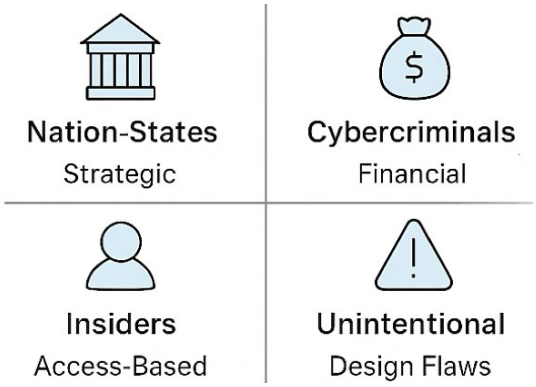

Figure 3: Key threat actors and their primary motivations in exploiting AI systems.

Each actor type contributes to the multifaceted threat landscape surrounding AI systems. Their motivations, ranging from geopolitical dominance to financial gain or negligence, must be considered when designing threat models and mitigation strategies.

6 Methodologies and Frameworks

Effective cybersecurity for AI systems requires structured methodologies that go beyond traditional IT risk management. These methodologies guide the identification, assessment, and mitigation of AI-specific threats, incorporating both technical and organizational dimensions. Methodologies and frameworks provide structured, often governance-oriented processes for identifying, assessing, and managing AI security risks across the system lifecycle. Several frameworks have emerged to address the unique challenges posed by AI:

- **NIST AI Risk Management Framework (AI RMF)** defines four core functions, Govern, Map, Measure, and Manage, to help organizations systematically address risks across the AI lifecycle [30]. For instance, an organization deploying an AI-based fraud detection system can use the 'Map' and 'Measure' functions to identify data-quality risks and evaluate model robustness before the system is moved into production.
- **ISO/IEC 23894:2023** complements general information security standards (e.g., ISO/IEC 27001) by introducing AI-specific risk considerations, including emergent behavior and data-centric vulnerabilities [31, 32].
- **OCTAVE and FAIR** provide strategic modeling and risk quantification tools. OCTAVE focuses on organizational context and asset prioritization, while FAIR enables probabilistic analysis of threat events and their impact [33, 34].
- **TARA (Threat Assessment and Remediation Analysis)**, developed by Intel, offers a structured approach to identifying and prioritizing threats based on attack vectors and system exposure [35]. As an example, TARA can be used to prioritize threats to an AI-enabled authentication system by identifying high-impact attack vectors such as model inversion or credential spoofing, allowing security teams to focus mitigations where exposure is greatest.
- **Structured Assurance Cases (SACs)** are increasingly used in regulated environments to formally justify that AI systems meet defined safety and security goals. These cases support compliance and traceability in high-assurance domains [36].

These methodologies support a shift from reactive to proactive security management, enabling organizations to anticipate and mitigate risks before deployment. They also facilitate alignment with governance frameworks and stakeholder expectations.

7 Taxonomies and Classification Models

To effectively manage cybersecurity risks in AI systems, it is essential to apply structured taxonomies and classification models. These frameworks help categorize threats, vulnerabilities, and impacts, enabling systematic threat modeling and mitigation. Taxonomies and classification models serve as conceptual lenses for categorizing threats, vulnerabilities, and impacts, and are typically used within a larger methodology to structure analysis. Several established models have been adapted for AI-specific contexts:

- **STRIDE** categorizes threats into Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. In AI systems, it is used to identify vulnerabilities such as adversarial inputs, data tampering, and model extraction [28]. For example, applying STRIDE to an image-classification pipeline may reveal tampering risks where adversaries manipulate input data, or spoofing risks where attackers impersonate legitimate data sources to influence model behavior.
- The **CIA Triad**, Confidentiality, Integrity, and Availability, remains foundational for assessing AI risks. It is particularly relevant for evaluating threats like data leakage, model corruption, and service disruption [37].
- **LINDDUN** focuses on privacy threat modeling, especially for AI systems handling sensitive personal or behavioral data. It supports the identification of risks related to linkability, identifiability, and non-repudiation [38].
- The **OWASP Machine Learning Security Top 10** provides a curated list of critical threats to AI systems, including adversarial inputs, insecure data pipelines, model theft, and poisoning attacks [39]. A concrete example is model inversion, where an attacker queries a deployed model to reconstruct sensitive information about individuals in the training data, a threat highlighted explicitly in OWASP's ML security guidance [5].
- **MITRE ATT&CK for ML** and **MITRE ATLAS** extend traditional adversarial frameworks to cover AI-specific tactics and techniques. These include model inversion, evasion, and poisoning, offering detailed mappings of attack paths and countermeasures [27].
- **AI-adapted attack trees** and **red teaming frameworks** simulate real-world adversarial scenarios, helping analysts trace potential compromise pathways and evaluate system resilience [40].

These classification models support both proactive and reactive security strategies. They enable de-

fenders to anticipate attack vectors, assess system exposure, and implement targeted mitigations aligned with AI system behavior and deployment context.

8 Tools for AI Security Analysis

A growing ecosystem of tools supports the identification, analysis, and mitigation of cybersecurity threats in AI systems. Tools provide practical and operational mechanisms that implement or reinforce the methodologies and taxonomies described in the previous sections. They include testing environments, documentation systems, red teaming platforms, and diagnostic utilities that assist practitioners in evaluating system robustness and addressing AI-specific vulnerabilities. Together, these tools span threat modeling, adversarial testing, explainability, and operational monitoring, and they enable both proactive and reactive security strategies.

- **Threat modeling tools** such as OWASP Threat Dragon and Microsoft's Threat Modeling Tool assist in building structured threat maps using frameworks like STRIDE. These tools help visualize attack surfaces and prioritize mitigation efforts [39, 41].
- **Adversarial testing environments** like IBM's Adversarial Robustness Toolbox (ART) and SecML allow researchers and practitioners to simulate adversarial scenarios and evaluate model resilience. These platforms support robustness testing against evasion, poisoning, and inversion attacks [42, 43]. For example, ART can be used to generate adversarial perturbations against image-classification models to assess whether small input modifications can force misclassification, enabling security teams to evaluate evasion susceptibility under realistic attack conditions.
- **Explainability and documentation tools** support transparency and governance. Google's Model Card Toolkit standardizes model reporting, while Microsoft's Responsible AI Dashboard adds bias detection, fairness metrics, and error analysis [44, 41]. For example, a model card for an AI-based credit-scoring system can document training data, performance limits, and fairness metrics for compliance. Teams should maintain model cards, enable post-deployment explainability, and use continuous monitoring to ensure transparency.
- **Red teaming frameworks** and **AI-adapted attack trees** simulate real-world adversarial behavior, helping organizations identify unknown vulnerabilities and test system robustness under realistic conditions [40].
- **Structured Assurance Cases (SACs)** provide formal documentation of how AI systems meet safety, security, and compliance goals, particularly in regulated domains such as healthcare, finance, and defense [36].

These tools support a lifecycle approach to AI security by enabling threat anticipation during design, resilience testing during development, and continuous monitoring during deployment. They are increasingly used in industry, although with varying applicability. Microsoft and IBM platforms integrate naturally into enterprise environments, while Google's Model Card Toolkit is widely used for documentation and is often complemented with sector-specific security solutions in regulated domains.

9 Stakeholders and Governance

Securing AI systems against cybersecurity threats requires coordinated efforts across technical, regulatory, and policy domains. A diverse set of stakeholders contribute to the development of standards, frameworks, and best practices for AI security governance.

- **NIST (USA)** plays a central role in defining AI risk management methodologies, including the AI RMF and the broader Cybersecurity Framework [45]. For example, organizations adopting the NIST AI RMF often use it to align their internal model development and monitoring practices with structured governance requirements, helping ensure that AI systems meet defined security and reliability criteria [10].
- **ISO/IEC (Global)** provides technical standards for AI lifecycle management, system robustness, and information security. ISO/IEC 23894:2023 specifically addresses AI-related risks [31, 32].
- **ENISA (EU)** publishes threat taxonomies and guidance documents tailored to AI threat scenarios [46]. A practical example is ENISA's guidance being used by operators of critical infrastructure, such as energy networks, to assess AI-related vulnerabilities and align their risk assessments with EU-wide cybersecurity recommendations.

- **European Commission** enforces compliance through instruments such as the AI Act and GDPR [47].
- **OECD** contributes high-level governance principles and policy recommendations for trustworthy AI [48].

Supporting institutions include:

- **MITRE**, which develops adversarial modeling frameworks such as ATT&CK for ML and ATLAS [27].
- **IEEE**, which focuses on ethical AI standards, transparency, and bias mitigation [49].
- **CISA (USA)** and **ECCC (EU)** contribute through infrastructure protection, vulnerability alerts, and innovation support [50, 51].

These stakeholders collectively shape the governance ecosystem for AI cybersecurity. Their contributions span technical standards, regulatory enforcement, ethical guidance, and operational support, highlighting the need for multi-stakeholder collaboration to address the complex risks posed by AI systems.

10 Literature Search Approach

This section outlines the methodology used to identify and analyze relevant literature on cybersecurity threats to and from AI systems. A dual-source strategy was employed, combining academic and grey literature to ensure comprehensive coverage of technical, organizational, and policy dimensions.

Semi-Structured Academic Search

A semi-structured search was conducted using Google Scholar, targeting peer-reviewed publications from 2021–2025. Seven search strings were formulated to capture the dual role of AI in cybersecurity:

- Cybersecurity risks to and from AI systems
- AI in cyberattacks and cyber defense
- Adversarial attacks and model vulnerabilities
- AI-enabled threats and generative attacks
- Broad phrase-based discovery
- AI in critical infrastructure contexts
- Robust and secure AI model development

Inclusion criteria required relevance to AI–cybersecurity intersections, credible sources, and technical or policy-level insights. Exclusion criteria filtered out ethics-only papers, unverifiable sources, duplicates, and outdated items.

Grey Literature Search

A complementary search targeted authoritative non-academic sources, including reports from NIST, OWASP, ENISA, OECD, IBM, and Deloitte. Grey literature refers to such non-academic sources, typically not peer-reviewed, including industry reports, policy briefs, technical documentation, and standards published by public agencies and private organizations. These materials often provide practical insights and early signals that complement academic research.

To ensure consistency, the same balanced query logic used in the academic search was applied. These grey literature sources contributed practical and policy-level perspectives that frequently precede formal academic publication, offering a broader understanding of the cybersecurity implications of AI systems.

Summary of Findings

The combined search yielded a total of 212 unique sources, consisting of both academic and grey literature. These results are summarized in Table 1, which outlines the initial records, filtering steps, and final retained items for each source type.

Table 1: Summary of Literature Search Results

Source Type	Initial Records	After Filtering	Final Retained
Academic (Google Scholar)	~300	286 unique	154
Grey Literature	85	65 unique	58
Total Retained	—	—	**212**

This consolidated dataset forms the basis for the thematic analysis presented in the following chapters, enabling a structured understanding of cybersecurity threats related to AI systems.

11 Search Results

This section outlines key patterns from both academic and grey literature searches. The academic search spans years, platforms, and topics, while the grey literature offers focused insights from industry and policy. Together, they reflect emerging priorities in AI-related cybersecurity.

Analysis of Academic Search Results

The academic search yielded 154 references, revealing clear temporal and thematic trends. Figure 4 shows the yearly distribution of academic references, highlighting a sharp increase in 2024.

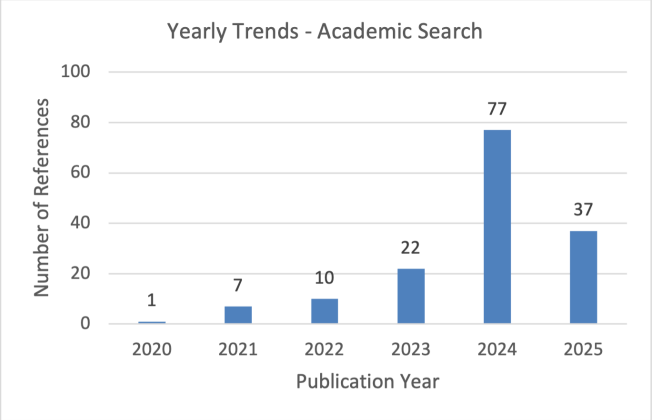

Figure 4: Yearly distribution of academic references.

Publications surged in 2024 (77 references), reflecting heightened attention to cybersecurity concerns and emerging AI-related themes. The lower count for 2025 (37 references) likely reflects the timing of the search, as many 2025 publications may not yet have been released or indexed (mid June).
Figure 5 shows academic topic trends across years. Cyber Risk & Defense and Threat Detection dominate the thematic landscape, with sharp increases in 2024.

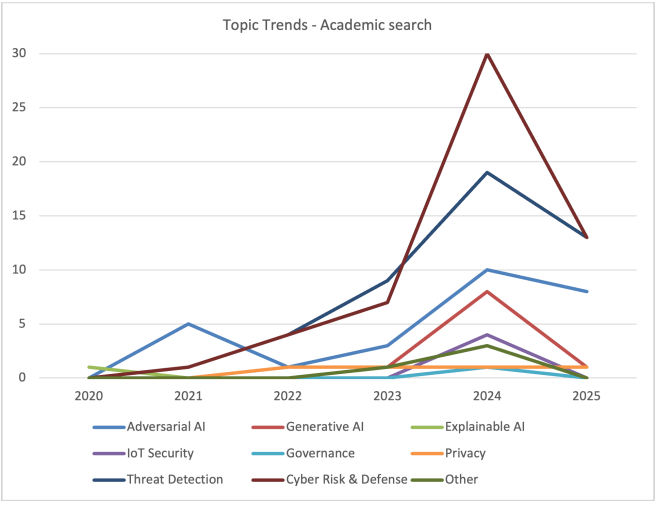

Figure 5: Academic topic trends across years.

Adversarial AI also shows strong presence, while Generative AI rises in 2024 but remains secondary compared to these leading themes. Governance and Privacy emerge as secondary but growing topics. Explainable AI and IoT Security remain relatively stable and less represented. Generative AI growth is likely underrepresented in these metrics because it often serves as a supporting tool or technique across other topics, and its sharp rise from 2024 onward suggests an accelerating influence on the overall thematic landscape.
Figure 6 presents the source breakdown, with ResearchGate as the leading platform.

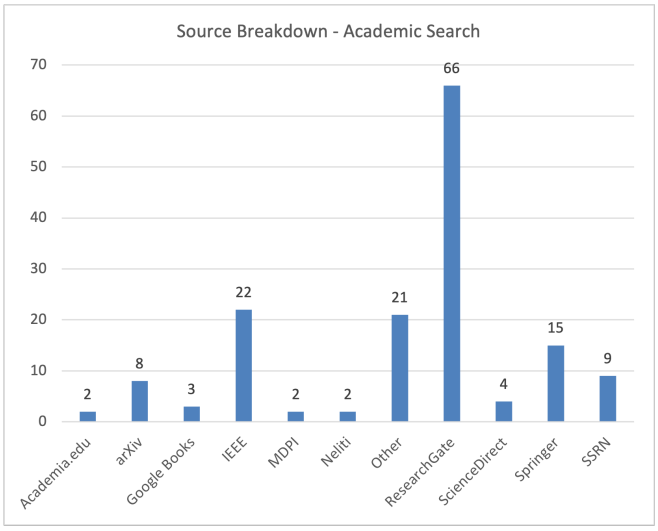

Figure 6: Source breakdown of academic references.

ResearchGate accounts for the largest share (66 references), followed by IEEE (22) and Springer (15). Preprint platforms such as arXiv (8) and SSRN (9) indicate a strong pre-publication culture. Other sources collectively contribute 34 references, reflecting diversity in dissemination.

Analysis of Grey Literature Search Results

The grey literature search yielded 58 references, revealing clear temporal and thematic trends. Figure 7 shows the yearly distribution of grey literature references, highlighting a sharp increase in 2024.

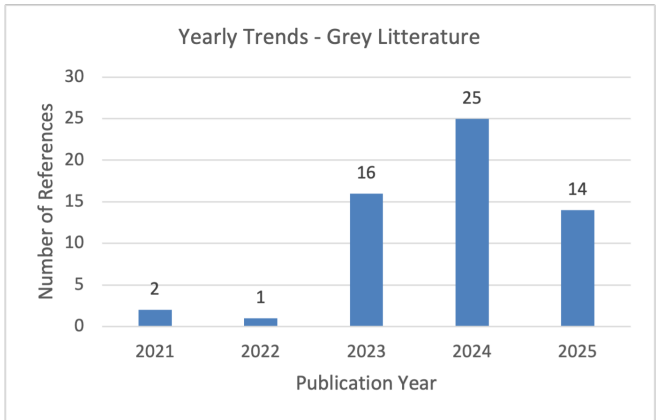

Figure 7: Yearly distribution of grey literature references.

Publications surged in 2024 (25 references), reflecting heightened industry and policy attention to AI-related cybersecurity risks. The lower count for 2025 (14 references) reflects that the search was conducted before the end of the year (late October), and the dataset therefore only includes publications indexed up to that point. Additional references are likely to appear as the year progresses.

Figure 8 illustrates topic trends across grey literature. Threat Detection dominates in 2024 and remains significant in 2025.

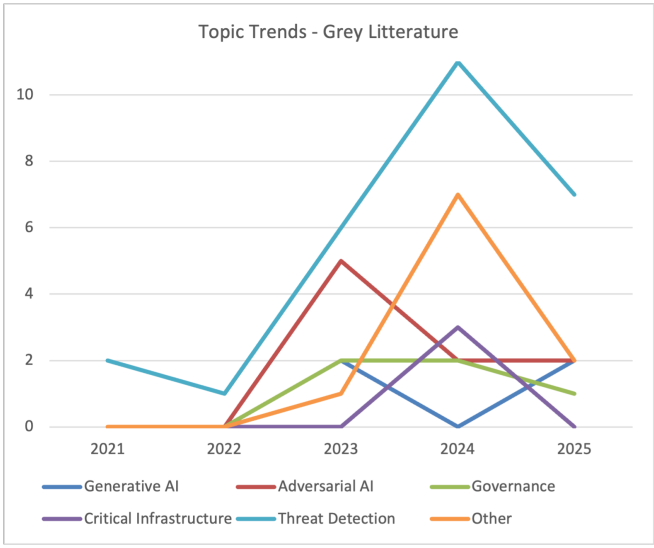

Figure 8: Grey literature topic trends across years.

Adversarial AI maintains a moderate presence across the period, while Generative AI is limited in absolute counts. Governance and Critical Infrastructure appear consistently, though at modest levels. Generative AI's influence is likely understated in these metrics because it often supports other areas, and its emerging role suggests potential for significant growth in future grey literature.

Figure 9 presents the source breakdown, grouped by publication type.

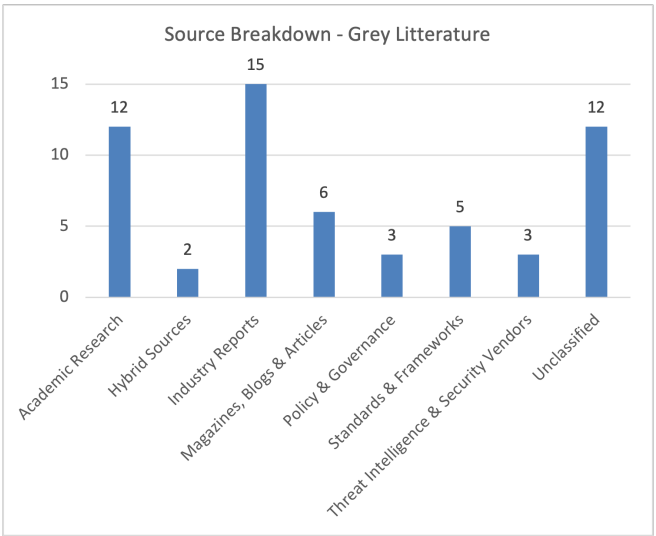

Figure 9: Source breakdown of grey literature references.

Industry reports account for the largest share (15 references), followed by Academic Research (12) and Unclassified sources (12). Other categories include Magazines, Blogs & Articles (6), Standards & Frameworks (5), Threat Intelligence & Security Vendors (3), Policy & Governance (3), and Hybrid Sources (2). This distribution reflects broad engagement across industry, academia, and policy domains, with Unclassified sources indicating diverse contributions beyond formal categories.

While academic sources emphasize technical mechanisms such as adversarial robustness and model inversion, grey literature tends to highlight operational concerns, governance, and real-world deployment risks. Explainable AI and IoT security appeared less frequently, suggesting potential gaps in current research and industry focus despite their relevance to transparency and edge deployment.

Limitations of the Literature Search

The academic and grey literature searches in this report were conducted using structured queries and filtering criteria, with the following limitations. The search was finalized in mid June and late October 2025, which means recent publications may be underrepresented, especially for the current year. The academic search relied primarily on Google Scholar, which may favor open-access platforms and preprints. The grey literature search focused on major organizations and vendors, potentially overlooking smaller or regional sources. Additionally, the phrasing of search strings and exclusion of ethics-only papers may have limited the scope of interdisciplinary insights. These factors affect the scope and balance of the findings.

Summary of Search Results

The combined literature search yielded 212 retained sources: 154 academic and 58 grey literature references. Academic publications showed a sharp increase in 2024, driven by interest in generative AI, adversarial attacks, and cybersecurity applications. Grey literature followed a similar trend, with industry and policy reports emphasizing threat detection, governance, and critical infrastructure protection.

Generative AI emerged as a rapidly growing theme across both datasets, alongside adversarial AI, cyber risk and defense, and governance, and is expected to influence other areas as a supporting technique. Explainability and privacy were present but less prominent. Academic sources were primarily retrieved from platforms like ResearchGate, IEEE, and Springer, while grey literature was drawn

from organizations such as NIST, ENISA, OWASP, and IBM.
Together, the results reflect a convergence of technical and strategic concerns, with growing attention to AI-specific vulnerabilities, threat modeling frameworks, and governance mechanisms. For complete reference lists, see Appendix A and Appendix B.

12 Conclusion

The integration of artificial intelligence into critical systems introduces a dual cybersecurity challenge: protecting AI systems from exploitation and preventing their misuse as threat vectors. This report has outlined the state of the art in both domains, highlighting emergent vulnerabilities such as adversarial attacks, data poisoning, and model inversion, as well as offensive capabilities including automated phishing, deepfake generation, and adaptive malware.

Key insights include:

- AI systems need specialized security strategies that go beyond traditional IT defenses. These include adversarial robustness, explainability, and continuous monitoring.
- The dual role of AI, as both target and tool, complicates threat modeling and necessitates integrated approaches to risk management.
- A wide range of actors, from nation-states to insiders, exploit AI systems for strategic, financial, or disruptive purposes. Their motivations must inform mitigation strategies.
- Methodologies such as NIST AI RMF, ISO/IEC 23894, FAIR, and SACs provide structured approaches to AI risk governance.
- Taxonomies like STRIDE, OWASP ML Top 10, and MITRE ATLAS support systematic threat classification and adversarial scenario modeling.
- Tools for threat modeling, adversarial testing, and explainability are essential for securing AI across its lifecycle.
- Multi-stakeholder governance, spanning technical standards, regulatory enforcement, and ethical oversight, is critical to addressing the complex risks posed by AI technologies.
- Securing AI requires a lifecycle approach: start with resilience by design (threat modeling, assurance cases, robustness testing) and extend to deployment transparency and continuous monitoring.

Looking ahead, emerging intersections between AI and quantum computing, as well as the rise of autonomous agents, may introduce new threat surfaces. Continued research is needed to anticipate these developments and adapt existing frameworks accordingly. The literature search reinforces these findings, revealing a strong academic focus on generative and adversarial AI, and complementary grey literature insights from industry and policy actors. Together, they highlight a convergence of technical innovation and strategic concern, underscoring the urgency of coordinated responses across sectors.

In conclusion, securing AI systems demands a multidisciplinary effort that combines cybersecurity expertise, machine learning robustness, and responsible governance. Future work should focus on harmonizing standards, improving tool interoperability, and fostering international collaboration to ensure that AI technologies remain trustworthy and resilient. This includes aligning technical safeguards with legal and ethical standards across jurisdictions, that is, across different regulatory systems, national laws, and sector-specific governance frameworks.

Acknowledgments

This research was supported by the CitCom.ai Testing and Experimentation Facility (TEF), co-funded by the European Union (`https://citcomtef.eu`). Additional support was provided by the STRIDE project (Secure and Resilient Infrastructure for Digital Enterprises), funded by Vinnova under reference number 2024-03263. The Center for Cybersecurity at RISE also contributed to this work via the CyRA project. Furthermore, we gratefully acknowledge the reviewers Saad Azar and David Eklund for their constructive feedback, which significantly improved the clarity and quality of this work.

Declaration of AI Assistance

The authors acknowledge the use of artificial intelligence (AI) tools during the preparation of this paper. Microsoft Copilot (based on GPT-4 and GPT-5) was used to assist with language refinement, structural suggestions, checking presence of references, and formatting. In addition, Copilot supported the conversion of references from Word to LaTeX format and assisted in the reference analysis by generating Excel files for further processing. Conceptual contributions, research design, data analysis, and interpretations were made by the authors. The AI tool did not replace human judgment or responsibility in the research process.

References

[1] J. Andress, *The basics of information security: Understanding the fundamentals of InfoSec in theory and practice*, 2nd ed. Syngress Media. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/C20130186424

[2] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," pp. 19–35. [Online]. Available: https://api.semanticscholar.org/CorpusID:4551073

[3] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318302565

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 3–18. [Online]. Available: http://dx.doi.org/10.1109/SP.2017.41

[5] Owasp. OWASP Machine Learning Security Top Ten 2023. [Online]. Available: https://owasp.org/www-project-machine-learning-security-top-10/docs/ML03_2023-Model_Inversion_Attack

[6] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. Allen, J. Steinhardt, C. Flynn, S. O'Heigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei, *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Apollo - University of Cambridge Repository. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/275332

[7] N. Papernot, P. Mcdaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:1090603

[8] H. Gentile. Bring light to the black box. [Online]. Available: https://www.ibm.com/think/insights/bring-light-to-the-black-box

[9] W. Isaac and J. Reno, "Ai product security: A primer for developers," *arXiv preprint arXiv:2304.11087*, 2023. [Online]. Available: https://arxiv.org/abs/2304.11087

[10] A. Vassilev, "Adversarial machine learning:: A taxonomy and terminology of attacks and mitigations," National Institute of Standards and Technology, resreport. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2.pdf

[11] O. M. Ijiga, I. P. Idoko, G. I. Ebiega, F. I. Olajide, T. I. Olatunde, and C. Ukaegbu, "Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention," vol. 11, pp. 001–004. [Online]. Available: https://oarjst.com/content/harnessing-adversarial-machine-learning-advanced-threat-detection-ai-driven-strategies

[12] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. ACM, pp. 1322–1333. [Online]. Available: https://dl.acm.org/doi/10.1145/2810103.2813677

[13] S. Shahriar, S. Allana, S. M. Hazratifard, and R. Dara, “A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle,” vol. 11, pp. 61 829–61 854. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2023.3287195

[14] I. H. Sarker, *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability*. Springer Nature Switzerland. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-54497-2

[15] M. K. Pratt. Emerging cyber threats in 2023 from AI to quantum to data poisoning. [Online]. Available: https://www.csoonline.com/article/651125/emerging-cyber-threats-in-2023-from-ai-to-quantum-to-data-poisoning.html

[16] M. Schmitt and I. Flechais, “Digital deception: generative artificial intelligence in social engineering and phishing,” vol. 57. [Online]. Available: http://dx.doi.org/10.1007/s10462-024-10973-2

[17] M. I. Khan, A. Arif, A. R. A. Khan, N. Anjum, and H. Arif, “The dual role of artificial Intelligence in cybersecurity: Enhancing defense and navigating challenges,” vol. 13, pp. 62–67. [Online]. Available: http://dx.doi.org/10.55524/ijircst.2025.13.1.9

[18] H. K. Pedarla, “The rise of AI-generated malware: Detection challenges and countermeasures,” vol. 11. [Online]. Available: https://www.ijirct.org/download.php?a_pid=2510016

[19] K. Jabbarova, “AI and cybersecurity - new threats and opportunities,” vol. 22, pp. 5955–5966, https://journlra.org/index.php/jra/article/view/754. [Online]. Available: https://www.pjlss.edu.pk/pdf_files/2024_2/9966-9975.pdf

[20] S. Miller, “Lethal autonomous weapon systems (LAWS): meaningful human Control, collective moral responsibility and institutional design,” vol. 27. [Online]. Available: http://dx.doi.org/10.1007/s10676-025-09874-x

[21] R. Crootof, “The killer robots are here: Legal and policy implications,” vol. 36, p. 1837. [Online]. Available: https://api.semanticscholar.org/CorpusID:110021524

[22] S. Feldstein, “The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression,” vol. 30, pp. 40–52. [Online]. Available: https://muse.jhu.edu/article/713721

[23] F. House, “Freedom on the Net 2023: The Repressive Power of Artificial Intelligence,” Freedom House, resreport. [Online]. Available: https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[24] W. Hoffman, “AI and the Future of Cyber Competition.” [Online]. Available: https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/

[25] M. Malatji, “Offensive artificial intelligence: Current state of the art and future directions,” in *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*. IEEE, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICDATE58146.2023.10248780

[26] H. Xinbo and L. Guo, “Navigating the risks of Generative AI: A comparative analysis of international regulatory approaches,” vol. 20, pp. 11–20. [Online]. Available: https://ph.pollub.pl/index.php/preko/article/view/7444

[27] MITRE Corporation, “MITRE ATLAS™,” accessed 2025. [Online]. Available: https://atlas.mitre.org

[28] A. Shostack, *Threat modeling: Designing for security*. John Wiley & Sons, accessed 2025. [Online]. Available: https://ieeexplore.ieee.org/book/9932141

[29] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “A survey on adversarial attacks and defences,” vol. 6, pp. 25–45. [Online]. Available: http://dx.doi.org/10.1109/TAI.2021.3056507

[30] National Institute of Standards and Technology, “AI Risk Management Framework.” [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[31] *ISO/IEC 23894:2023(en) Information technology — Artificial intelligence — Guidance on risk management*, Std. ISO/IEC 23 894:2023. [Online]. Available: https://www.iso.org/standard/77304.html

[32] A. Simonetta and M. C. Paoletti, “ISO/IEC Standards and Design of an Artificial Intelligence System,” in *CEUR Workshop Proceedings*. ceur-ws.org, pp. 39–43. [Online]. Available: https://ceur-ws.org/Vol-3916/paper_07.pdf

[33] B. Tucker, “OCTAVEÂ® FORTE and FAIR Connect Cyber Risk Practitioners with the Boardroom.” [Online]. Available: https://www.sei.cmu.edu/blog/octavea-forte-and-fair-connect-cyber-risk-practitioners-with-the-boardroom/

[34] J. B. Copeland. A FAIR Artificial Intelligence (AI) Cyber Risk Playbook. Accessed 2025. [Online]. Available: https://www.fairinstitute.org/blog/fair-artificial-intelligence-ai-cyber-risk-playbook

[35] I. C.-W. Contributors. Threat Agent Risk Assessment (TARA). [Online]. Available: https://cio-wiki.org/wiki/Threat_Agent_Risk_Assessment_(TARA)

[36] Society for Risk Analysis. The Society for Risk Analysis: A Community for Risk Science. [Online]. Available: https://www.sra.org/

[37] ISO/IEV, *ISO/IEC 27000:2018 Information technology — Security techniques — Information security management systems — Overview and vocabulary*, Std. [Online]. Available: https://www.iso.org/standard/73906.html

[38] W. Kim. LINDDUN Threat Modeling. [Online]. Available: https://linddun.org/

[39] OWASP. OWASP Machine Learning Security Top Ten. [Online]. Available: https://owasp.org/www-project-machine-learning-security-top-10/

[40] OpenAI. Security & privacy. [Online]. Available: https://openai.com/security-and-privacy

[41] Microsoft Corporation. Microsoft Threat Modeling Tool. [Online]. Available: https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool

[42] M.-I. Nicolae, “Welcome to the Adversarial Robustness Toolbox — Adversarial Robustness Toolbox 1.17.0 documentation.” [Online]. Available: https://adversarial-robustness-toolbox.readthedocs.io/en/latest/?utm_source=chatgpt.com

[43] M. Melis, “SecML: Secure and Explainable Machine Learning in Python — SecML 0.15 documentation.” [Online]. Available: https://secml.readthedocs.io/en/v0.15/

[44] Google, “model-card-toolkit: A toolkit that streamlines and automates the generation of model cards.” [Online]. Available: https://github.com/tensorflow/model-card-toolkit

[45] National Institute of Standards and Technology. National Institute of Standards and Technology. U.S. Department of Commerce. [Online]. Available: https://www.nist.gov/

[46] ENISA. European Union Agency for Cybersecurity. [Online]. Available: https://www.enisa.europa.eu

[47] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[48] OECD. Organisation for Economic Co-operation and Development. [Online]. Available: https://www.oecd.org/

[49] IEEE. Autonomous and Intelligent Systems (AIS). [Online]. Available: https://standards.ieee.org/initiatives/autonomous-intelligence-systems/

[50] CISA. America's Cyber Defense Agency. [Online]. Available: https://www.cisa.gov

[51] ECCC. European Cybersecurity Competence Centre and Network. [Online]. Available: https://cybersecurity-centre.europa.eu/index_en

Appendix A: Academic References

1. Perumal, A. P., Chintale, P., & Molleti, R. (2024). Risk Assessment of Artificial Intelligence Systems in Cybersecurity. American Journal of Information Science.
2. Emehin, O., Akanbi, I., Emeteveke, I., & Adeyeye, O. J. (2024). Enhancing Cybersecurity with Safe and Reliable AI: Mitigating Threats While Ensuring Privacy Protection. SSRN Electronic Journal.
3. Gilbert, C., & Gilbert, M. (2024). The Impact of AI on Cybersecurity Defense Mechanisms: Future Trends and Challenges. SSRN Electronic Journal.
4. Novaes Neto, N., & Pearlson, K. (2024). Understanding the Cyber Risks of Artificial Intelligence: An Ongoing, Comprehensive, Multi-Faceted Approach for CIOs, CTOs and CSOs. SSRN Electronic Journal.
5. Farooq, M., & Zaffar, F. (2024). AI in Cybersecurity: Enhancing Threat Detection. AlgoVista Journal of AI and Computer Science.
6. Priyadharshini, S. L., Abbas, R., Arafat, Y., & Batool, W. (2024). Cybersecurity in AI-Driven IT Environments: A Study on Vulnerabilities and Mitigation Strategies. ResearchGate.
7. Chettier, T. M., Boyina, V. A. K., & Rangineni, S. (2024). AI-Powered Risk Assessment and Compliance in Cloud Cybersecurity. ResearchGate.
8. Sood, A. (2024). Combating Cyberattacks Targeting the AI Ecosystem: Assessing Threats, Risks, and Vulnerabilities. Google Books.
9. Bashynska, I., & Prokopenko, O. (2024). Mitigating Cyber Risks in AI-Driven Circular Economy Implementations. Scientific Journal of Bielsko-Biala School of Finance and Law.
10. Rao, P. S., Krishna, T. G., & Mahboub, M. A. (2024). AI in Cybersecurity: Challenges, Directions, and Research Needs—A Review. International Research Journal.
11. Siddique, R. A., & Jahankhani, H. (2024). An Investigation of Artificial Intelligence (AI) and Cybersecurity: Case of AI Integration in German Cybersecurity Strategy. Springer.
12. Malatji, M., & Tolah, A. (2024). Artificial Intelligence (AI) Cybersecurity Dimensions: A Comprehensive Framework for Understanding Adversarial and Offensive AI. AI and Ethics.
13. Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Navigating AI Cybersecurity: Evolving Landscape and Challenges. Journal of Intelligent Learning Systems.
14. Sharma, D. P., Lashkari, A. H., Firoozjaei, M. D., & Mahdavifar, S. (2025). Understanding AI in Cybersecurity and Secure AI. In Progress in IS. Springer.
15. Umeh, I. I. (2025). Enhancing Cybersecurity in the Age of AI: Challenges and Solutions. Academia.edu.
16. JothiShri, S., Upender, T., & Ravikumar, R. J. (2024). AI Cyber Security: Enhancing Network Security with Deep Learning for Real-Time Threat Detection and Performance Evaluation. IEEE.
17. Minhaj, S. M. U. H. (2023). Study of Artificial Intelligence in Cyber Security and the Emerging Threat of AI-Driven Cyber Attacks and Challenge. SSRN Electronic Journal.

18. Tumma, C. (2024). AI-Driven Cybersecurity Solutions for Enhancing IoT Network Security: A Comprehensive Approach. ResearchGate.

19. Kolosnjaji, B., Xiao, H., Xu, P., & Zarras, A. (2024). Artificial Intelligence for Cybersecurity: Develop AI Approaches to Solve Cybersecurity Problems in Your Organization. Google Books.

20. Nacheva, R., & Azeroual, O. (2024). Security of AI-Powered Systems: Threat Intelligence on the Edge. 8th International Symposium on Computational Intelligence. IEEE.

21. Sarma, W., Srivastava, A., & Sresth, V. (2024). AI-Driven Cybersecurity for IoT Ecosystems: Leveraging Machine Learning for Proactive Threat Detection and Autonomous Defense Mechanisms. ResearchGate.

22. Romanous, E., & Ginger, J. (2024). AI Efficiency in Cybersecurity: Estimating Token Consumption for Optimal Operations. IEEE.

23. Goffer, M. A., Uddin, M. S., & Hasan, S. N. (2025). AI-Enhanced Cyber Threat Detection and Response: Advancing National Security in Critical Infrastructure. ResearchGate.

24. Tiwari, S., Sresth, V., & Srivastava, A. (2020). The Role of Explainable AI in Cybersecurity: Addressing Transparency Challenges in Autonomous Defense Systems. International Journal of Computer Science.

25. Metta, S., Chang, I., Parker, J., & Roman, M. P. (2024). Generative AI in Cybersecurity. arXiv preprint arXiv.

26. Sharma, D. P. (2024). Understanding AI in Cybersecurity and Secure AI: Challenges, Strategies and Trends. Google Books.

27. Saxena, R., Baskar, A., Haroon, S., & Hayat, S. (2024). Cybersecurity Concerns of Artificial Intelligence Applications on High-Performance Computing Systems. ResearchGate.

28. Ee, S., O'Brien, J., Williams, Z., & El-Dakhakhni, A. (2024). Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-in-Depth Approach. arXiv.

29. Hamon, R., Junklewitz, H., Garrido, J. S., & Sanchez, I. (2024). Three Challenges to Secure AI Systems in the Context of AI Regulations. IEEE Access.

30. Abid, N. (2024). Cybercrime and Cybersecurity in the Age of AI: Exploring the Challenges and Opportunities Presented by ChatGPT. Global Journal of Multidisciplinary Sciences and Arts.

31. Bogdanov, D., Etti, P., & Kamm, L. (2024). Artificial Intelligence System Risk Management Methodology Based on Generalized Blueprints. IEEE.

32. Sobien, D., Yardimci, M. O., Nguyen, M. B. T., & Mao, W. Y. (2023). AI for Cyberbiosecurity in Water Systems—A Survey. Springer.

33. Aarush, I. C., & El Saadawi, N. (2024). Cybersecurity Governance Models for AI-Driven Enterprises. ResearchGate.

34. Neoaz, N., Bacha, A., Khan, M., & Sherani, A. M. K. (2025). AI in Motion: Securing the Future of Healthcare and Mobility through Cybersecurity. Asian Journal of Emerging Sciences and Humanities.

35. Sood, A. K., Zeadally, S., & Hong, E. K. (2025). The Paradigm of Hallucinations in AI-Driven Cybersecurity Systems: Understanding Taxonomy, Classification Outcomes, and Mitigations. Computers and Electrical Engineering.

36. Gopireddy, R. R. (2024). Securing AI Systems: Protecting Against Adversarial Attacks and Data Poisoning. Journal of Scientific and Engineering Research.

37. Raji, A. N., Olawore, A. O., Ayodeji, A., & Joseph, J. (2023). Integrating Artificial Intelligence, Machine Learning, and Data Analytics in Cybersecurity. ResearchGate.

38. Chawande, S. (2024). Insider Threats in Highly Automated Cyber Systems. ResearchGate.

39. Adeyeye, O. J., Akanbi, I., & Emeteveke, I. (2024). Leveraging Secured AI-Driven Data Analytics for Cybersecurity: Safeguarding Information and Enhancing Threat Detection. International Journal of Cybersecurity.

40. Rampášek, M., Mesarčík, M., & Andraško, J. (2025). Evolving Cybersecurity of AI-Featured Digital Products and Services: Rise of Standardisation and Certification?. Computer Law & Security Review.

41. Walter, H. (2024). AI for Cyber Defense: Leveraging Machine Learning to Detect and Prevent Threats. ResearchGate.

42. Lohn, A. J. (2025). The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict. arXiv preprint arXiv:2504.13371.

43. Chakraborty, A., Biswas, A., & Khan, A. K. (2023). Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation. In AI for Societal Progress. Springer.

44. George, A. S. (2024). Riding the AI Waves: An Analysis of Artificial Intelligence's Evolving Role in Combating Cyber Threats. Partners Universal International Innovation Journal.

45. Ndlovu, M. P., & Tsibolane, P. (2025). Exploring Organizational Resilience Towards AI-Driven Cyber Threats: A Systematic Literature Review. ResearchGate.

46. Ijiga, O. M., Idoko, I. P., Ebiega, G. I., & Olajide, F. I. (2024). Harnessing Adversarial Machine Learning for Advanced Threat Detection: AI-Driven Strategies in Cybersecurity Risk Assessment and Fraud Prevention. Journal of Scientific Research.

47. Kodhai, E., & Harishwar, B. (2025). Emerging Threats to Personal Data: AI-Powered Cyberattacks. IEEE Conference on Data Science, Agents & Artificial Intelligence.

48. Agarwal, G. (2025). Explainable AI (XAI) for Cyber Defense: Enhancing Transparency and Trust in AI-Driven Security Solutions.* International Journal of Advanced Research in Science.

49. Awotidebe, M. (2025). The Rise of Intelligent Threats: Exploring AI-Driven Cybercrime in the Digital Era. ResearchGate.

50. Haryanto, C. Y., Elvira, A. M., Nguyen, T. D., et al. (2024). Contextualized AI for Cyber Defense: An Automated Survey Using LLMs. IEEE Conference on Security of Information.

51. Balan, M. (2022). AI-Powered IAM and Threat Intelligence: Safeguarding Patient Data in the Age of Cybersecurity Breaches. ResearchGate.

52. Sugumaran, D., John, Y. M. M., Joshi, K., et al. (2023). Cyber Defence Based on Artificial Intelligence and Neural Network Model in Cybersecurity. Eighth IEEE Conference Proceedings.

53. Singh, T. (2025). Artificial Intelligence-Driven Cyberattacks. In Cybersecurity, Psychology and People Hacking. Springer.

54. Bright, S. (2024). Smarter Hackers, Smarter Threats: AI's Role in Multi-Cloud Security Breaches. ResearchGate.

55. Guo, W., Potter, Y., Shi, T., Wang, Z., & Zhang, A. (2025). Frontier AI's Impact on the Cybersecurity Landscape. arXiv preprint arXiv:

56. Kurtović, H., Šabanović, E., & Almisreb, A. A. (2024). Exploring the Dark Side: A Systematic Review of Generative AI's Role in Network Attacks. Conference of Recent Trends, Springer.

57. Olaoye, G. (2025). AI-Driven Intrusion Detection and Prevention Systems (IDPS) for Cloud Security. SSRN Electronic Journal.

58. Duary, S., Choudhury, P., & Mishra, S. (2024). Cybersecurity Threats Detection in Intelligent Networks Using Predictive Analytics Approaches. IEEE Conference Proceedings.

59. Lebed, S. V., Namiot, D. E., Zubareva, E. V., & Khenkin, P. V. (2024). Large Language Models in Cyberattacks. Doklady Mathematics, Springer.

60. Ogiela, M. R., & Ogiela, L. (2024). AI-Based Cybersecurity Systems. In International Conference on Advanced Information Systems. Springer.

61. Aldhamer, M. (2023). The Impact of Artificial Intelligence on the Future of Cybersecurity. Multi-knowledge Electronic Comprehensive Journal for Education and Science Publications.

62. AbdelRahman, A. A. B., Abbas, H. H., et al. (2024). Decoding Personal Security—Strategies to Safeguard Humans in the Era of Intelligent Machines. Proceedings of the 36th Conference, IEEE.

63. Okika, N., Okoh, O. F., & Etuk, E. E. (2025). Mitigating Insider Threats and Social Engineering Tactics in Advanced Persistent Threat Operations Through Behavioral Analytics and Cybersecurity Training. International Journal of Advance Research.

64. Guduru, S. (2025). Autonomous Cyber Defense: LLM-Powered Incident Response with LangChain and SOAR Integration. ResearchGate.

65. Samonte, M. J. C., Goc-ong, A. E., et al. (2024). Evaluating the Effectiveness of Artificial Intelligence in Integrated System Architectures to Combat Cybersecurity Threats. IEEE 7th International Conference.

66. Senewirathna, N. (2024). Quantum Computing and Its Impact on Information Warfare—Threats and Cybersecurity Countermeasures. ResearchGate.

67. Girdhar, M., Hong, J., & Moore, J. (2023). Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology.

68. Olutimehin, A. T., Ajayi, A. J., Metibemu, O. C., et al. (2025). Adversarial Threats to AI-Driven Systems: Exploring the Attack Surface of Machine Learning Models and Countermeasures. SSRN Electronic Journal.

69. Shayea, G. G., Zabil, M. H. M., Habeeb, M. A., et al. (2025). Strategies for Protection Against Adversarial Attacks in AI Models: An In-Depth Review. Journal of Intelligent Systems.

70. Bhuiyan, S., & Park, J. S. (2025). Cybersecurity Threats and Mitigation Strategies in AI Applications. Journal of The Colloquium for Information Systems Security Education (CISSE).

71. Anthi, E., Williams, L., Rhode, M., Burnap, P., et al. (2021). Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems. Journal of Information Security and Applications, 59, 102812.

72. Syed, S. A. (2025). Adversarial AI and Cybersecurity: Defending Against AI-Powered Cyber Threats. Iconic Research and Engineering Journals.

73. Ghimire, S., & Thapaliya, S. (2024). AI-Driven Cybersecurity: Mitigating Prompt Injection Attacks Through Adversarial Machine Learning. NPRC Journal of Multidisciplinary Research.

74. Xu, J., Wang, Y., Chen, H., & Shen, Z. (2025). Adversarial Machine Learning in Cybersecurity: Attacks and Defenses. International Journal of Cybersecurity.

75. Mohammed, A. S., Jha, S., Tabbassum, A., et al. (2024). Assessing the Vulnerability of Machine Learning Models to Cyber Attacks and Developing Mitigation Strategies. IEEE Conference on Intelligent Systems.

76. Zou, J., Zhang, S., & Qiu, M. (2024). Different Attack and Defense Types for AI Cybersecurity. In International Conference on Knowledge Science. Springer.

77. Khaleel, Y. L., Habeeb, M. A., Albahri, A. S., et al. (2024). Network and Cybersecurity Applications of Defense in Adversarial Attacks: A State-of-the-Art Using Machine Learning and Deep Learning Methods. Journal of Intelligent Systems.

78. Mishra, R. (2021). Adversarial Attacks and Defenses in AI-Powered Cybersecurity Systems. Journal of Computing and Information Systems.

79. Priyadharshini, S. L., Abbas, R., Arafat, Y., Batool, W., et al. (2025). Cybersecurity in AI-Driven IT Environments: A Study on Vulnerabilities and Mitigation Strategies. ResearchGate.

80. Anny, D. (2025). Adversarial Attacks on Generative AI Models in Cloud Platforms: Detection and Mitigation Strategies. ResearchGate.

81. Chitimoju, S. (2024). Mitigating the Risks of Prompt Injection Attacks in AI-Powered Cybersecurity Systems. Journal of Computing and Information Systems.

82. Musser, M., Lohn, A., Dempsey, J. X., & Spring, J. (2023). Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications. arXiv Preprint.

83. Munir, M. (2025). Theoretical Approaches to Cybersecurity and Adversarial Attacks in AI-Driven Marketing: A Framework for Risk Mitigation. Journal of Digital Literacy and Learning.

84. Temitope, O., & Owoyemi, J. (2024). Developing a Deep Learning Framework for Detecting and Mitigating Adversarial Attacks on Generative AI Systems in Cybersecurity Applications. ResearchGate.

85. Lee, C., & Lee, S. (2023). Evaluating the Vulnerability of YOLOv5 to Adversarial Attacks for Enhanced Cybersecurity in MASS. Journal of Marine Science and Engineering, 11(4), 701.

86. Hoang, V. T., Ergu, Y. A., Nguyen, V. L., & Chang, R. G. (2024). Security Risks and Countermeasures of Adversarial Attacks on AI-Driven Applications in 6G Networks: A Survey. Journal of Network and Computer Applications.

87. Kiranbabu, M. N. V., & Viji, A. J. (2025). The Challenge of Adversarial Attacks on AI-Driven Cybersecurity Systems. Journal of Cybersecurity & Digital Trust.

88. Omoseebi, A. (2022). Adversarial Attacks on Cloud-Based AI Fraud Detection Models: Risks and Mitigation Strategies. ResearchGate.

89. Lin, J., Dang, L., Rahouti, M., & Xiong, K. (2021). ML Attack Models: Adversarial Attacks and Data Poisoning Attacks. arXiv preprint arXiv:2112.02797.

90. Viureanu, A., & Ionescu, B. (2024). AI Vulnerabilities. Database Journal.

91. Kuppa, A., & Le-Khac, N. A. (2021). Adversarial XAI Methods in Cybersecurity. IEEE Transactions on Information Forensics and Security, 16, 4567–4579.

92. Sajjadi, S. H., Karamizadeh, S., & Yadegari, A. (2024). Defending Against Adversarial Attacks in Artificial Intelligence Technologies. International Journal of ICT, 12(2).

93. Josyula, M. M., & Saidireddy, M. (2024). Mitigating Adversarial and AI-Evasion Attacks in Cybersecurity: Challenges and Solutions. Tuijin Jishu/Journal of Propulsion Technology.

94. Li, L. (2024). Comprehensive Survey on Adversarial Examples in Cybersecurity: Impacts, Challenges, and Mitigation Strategies. arXiv preprint arXiv:2412.12217.

95. Chitimoju, S. (2023). The Risks of AI-Generated Cyber Threats: How Language Models Can Be Weaponized for Attacks. International Journal of Digital Innovation.

96. Sharma, D. P., Habibi Lashkari, A., & Firoozjaei, M. D. (2025). AI in Security. In Applications of AI in Cybersecurity and Beyond. Springer.

97. Usman, Y., Upadhyay, A., Gyawali, P., et al. (2024). Is Generative AI the Next Tactical Cyber Weapon for Threat Actors? Unforeseen Implications of AI-Generated Cyber Attacks. arXiv preprint arXiv:2401.01234.

98. Danish, M., & Siraj, M. M. (2025). AI and Cybersecurity: Defending Data and Privacy in the Digital Age. Journal of Engineering and Computational Intelligence Research.

99. Shrestha, L., Balogun, H., & Khan, S. (2025). AI-Driven Phishing: Techniques, Threats, and Defence Strategies. In International Conference on Global Cybersecurity. Springer.

100. McCall, A. (2024). Cybersecurity in the Age of AI and IoT: Emerging Threats and Defense Strategies. ResearchGate.

101. Antipova, T., Riurean, S., Riurean, P., & Bolog, G. (2025). AI-Powered Tools: Threat, Defense, and Cyber-Resilience for Individuals. In Digital Technology Platforms for Secure Societies. Springer.

102. Pakina, A. K., Kejriwal, D., & Pujari, T. D. (2025). Adversarial AI in Social Engineering Attacks: Large-Scale Detection and Automated Countermeasures. International Journal of Science and Engineering Research.

103. Sharma, S., & Dutta, N. (2024). Examining ChatGPT's and Other Models' Potential to Improve the Security Environment Using Generative AI for Cybersecurity. ResearchGate.

104. Manoharan, A., & Sarker, M. (2023). Revolutionizing Cybersecurity: Unleashing the Power of Artificial Intelligence and Machine Learning for Next-Generation Threat Detection. DOI: https://doi.org/10.56726

105. Blake, H. (2025). AI-Powered Threats in Supply Chains: A Looming Cybersecurity Challenge. ResearchGate.

106. Bondhala, S. (2024). Cybersecurity in AI-Driven Data Centers: Reinventing Threat Detection. ResearchGate.

107. Mathew, A. (2023). Cybercrime-as-a-Service & AI-Enabled Threats. International Journal of Computer Science and Mobile Computing.

108. Sharma, G., Malley, D. J., Parle, D., et al. (2024). Analysis and Study of Intelligent Testbed for Safeguarding Nuclear and Defense Industry from AI-Enabled Cyberattacks. In Proceedings of the Conference on Technology and Security. IEEE.

109. Jena, J. (2024). Emerging Threats in Generative AI: Strategies for Safeguarding Against New Cyber Risks. ResearchGate.

110. George, C., & Dominion Heritage, D. D. (2024). The Role of Artificial Intelligence in Enhancing Cybersecurity. ResearchGate.

111. Alotaibi, L., Seher, S., et al. (2024). Cyberattacks Using ChatGPT: Exploring Malicious Content Generation Through Prompt Engineering. In ASU International Cybersecurity Conference. IEEE.

112. Shafique, R., Rustam, F., Choi, G. S., et al. (2024). In-Vehicle Networks Security Using Transfer Learning Approach Against AI-Generated Cyberattacks. In 35th Irish Conference on Cybersecurity (ICCS). IEEE.

113. Owolabi, I. O., Mbabie, C. K., & Obiri, J. C. (2024). AI-Driven Cybersecurity in FinTech and Cloud: Combating Evolving Threats with Intelligent Defense Mechanisms. International Journal of Cybersecurity.

114. Pum, M. (2022). The Rise of AI-Powered Cybercrime: Implications for Digital Security and Policy. ResearchGate.

115. Pasupuleti, V. (2021). AI-Based Multimedia Security in Combating Adversarial Attacks, Deepfakes, and Ethical Concerns. International Journal of African & Asian Studies (IJAAS).

116. Akande, B. (2022). The Rise of AI-Driven Cybercrime: How Artificial Intelligence is Empowering Hackers. ResearchGate.

117. Dhoni, P. S., & Kumar, R. (2023). Artificial Intelligence and Cybersecurity: Roles of Generative AI Entities, Companies, Agencies and Governments in Enhancing Cybersecurity. ResearchGate.

118. John, B. (2025). Adapting to Advanced Threats: Celery Trap's Approach to Combating AI-Generated Phishing Campaigns. ResearchGate.

119. Singh, A. (2025). Preventive Measures for Network Security Using Generative Artificial Intelligence. SSRN Electronic Journal.

120. Tambi, V. K., & Singh, N. (2024). Investigating ChatGPT's and Other Models' Potential to Advance the Security Environment Using Generative AI for Cybersecurity. ResearchGate.

121. Charfeddine, M., Kammoun, H. M., Hamdaoui, B., et al. (2024). ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications. IEEE Transactions on Dependable and Secure Computing.

122. Atmaca, U. I., Le, A. T., Epiphaniou, G., et al. (2024). Emerging Threats of AI Integration in the Space User Segment: A Reference Architecture and Attack Tree Analysis. In IEEE 10th Conference on Recent Advances in Space Technologies.

123. Kumar, A. (2023). Next-Generation Approaches to Detecting and Preventing AI-Generated Phishing Scams. Eastern European Journal for Multidisciplinary Research, 3(4).

124. Zandi, G. R., Yaacob, N. A., Tajuddin, M., et al. (2024). Artificial Intelligence and the Evolving Cybercrime Paradigm: Current Threats to Businesses. Journal of Information Technology Management, 16(1).

125. Munson, T., Tao, V., & Mohr, J. J. (2024). The Double-Edged Sword of Generative AI. CPA Journal.

126. Kaur, J., Hasan, S. N., Orthi, S. M., Miah, M. A., et al. (2023). Advanced Cyber Threats and Cybersecurity Innovation—Strategic Approaches and Emerging Solutions. Journal of Computer and Communications, 11(6).

127. Zandi, G. R., Yaacob, N. A., & Tajuddin, M. (2024). Artificial Intelligence and the Evolving Cybercrime Paradigm: Current Threats to Businesses. Journal of Information.

128. Adewusi, A. O., Okoli, U. I., Olorunsogo, T., et al. (2024). Artificial Intelligence in Cybersecurity: Protecting National Infrastructure (USA Case Study). World Journal of Advanced Research.

129. Adhikari, D., & Thapaliya, S. (2024). Explainable AI for Cyber Security: Interpretable Models for Malware Analysis and Network Intrusion Detection. NPRC Journal of Multidisciplinary Research.

130. Mehmood, A., Shafique, A., Alawida, M., & Khan, A. N. (2024). Advances and Vulnerabilities in Modern Cryptographic Techniques: A Comprehensive Survey on Cybersecurity in the Domain of Machine/Deep Learning and Quantum Technologies. IEEE Access.

131. Jimmy, F. (2021). Emerging Threats: The Latest Cybersecurity Risks and the Role of Artificial Intelligence in Enhancing Cybersecurity Defenses. Valley International Journal Digital Library.

132. Mahmud, F., Barikdar, C. R., & Hassan, J. (2025). AI-Driven Cybersecurity in IT Project Management: Enhancing Threat Detection and Risk Mitigation. Journal of Cybersecurity Research.

133. Camacho, N. G. (2024). The Role of AI in Cybersecurity: Addressing Threats in the Digital Age. Journal of Artificial Intelligence General Science.

134. Abisoye, A., Akerele, J. I., Odio, P. E., & Collins, A. (2025). Using AI and Machine Learning to Predict and Mitigate Cybersecurity Risks in Critical Infrastructure. International Journal of Cybersecurity.

135. Rangaraju, S. (2023). AI Sentry: Reinventing Cybersecurity Through Intelligent Threat Detection. EPH-International Journal of Science and Engineering.

136. Sontan, A. D., & Samuel, S. V. (2024). The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities. World Journal of Advanced Research and Reviews.

137. Ahmed, S., Ahmed, I., & Kamruzzaman, M. (2022). Cybersecurity Challenges in IT Infrastructure and Data Management: A Comprehensive Review of Threats, Mitigation Strategies, and Future Trends. Mainstream Journal of Information Security.

138. Ahsan, M., Nygard, K. E., & Gomes, R. (2022). Cybersecurity Threats and Their Mitigation Approaches Using Machine Learning—A Review. Journal of Cybersecurity and Privacy.

139. Shahana, A., Hasan, R., Farabi, S. F., & Akter, J. (2024). AI-Driven Cybersecurity: Balancing Advancements and Safeguards. Journal of Computer.

140. Lekkala, S., Avula, R., & Gurijala, P. (2022). Big Data and AI/ML in Threat Detection: A New Era of Cybersecurity. Journal of Artificial Intelligence and Big Data.

141. Mohammed, A. (2023). The Paradox of AI in Cybersecurity: Protector and Potential Exploiter. Baltic Journal of Engineering and Technology.

142. Rajendran, R. M., & Vyas, B. (2023). Cyber Security Threat and Its Prevention Through Artificial Intelligence Technology. International Journal for Multidisciplinary.

143. Manduva, V. C. (2023). Artificial Intelligence and Cloud Computing: The Role of AI in Enhancing Cybersecurity. International Journal of Acta Informatica.

144. Yaseen, A. (2023). AI-Driven Threat Detection and Response: A Paradigm Shift in Cybersecurity. International Journal of Information and Cybersecurity.

145. Kalla, D., Kuraku, S., & Samaah, F. (2023). Advantages, Disadvantages and Risks Associated with ChatGPT and AI on Cybersecurity. Journal of Emerging Technologies.

146. Dandamudi, S. R. P., Sajja, J., & Khanna, A. (2025). Leveraging Artificial Intelligence for Data Networking and Cybersecurity in the United States. International Journal of Artificial Intelligence and Big Data.

147. Carbone, J. N., & Crowder, J. A. (2021). Artificially Intelligent Cyber Security: Reducing Risk and Complexity. In Advances in Artificial Intelligence and Applied Cognitive Computing. Springer.

148. Obioha Val, O., Lawal, T., & Olaniyi, O. O. (2025). Investigating the Feasibility and Risks of Leveraging Artificial Intelligence and Open Source Intelligence to Manage Predictive Cyber Threat Models. SSRN Electronic Journal.

149. Alagappan, A., & Andrews, L. J. B. (2022). Cybersecurity Risks Mitigation in the Internet of Things. In Proceedings of the 2nd International Conference on Cybersecurity. IEEE.

150. Adewusi, A. O., Chiekezie, N. R., & Eyo-Udo, N. L. (2022). The Role of AI in Enhancing Cybersecurity for Smart Farms. World Journal of Advanced Research and Reviews.

151. Shahriar, S., Allana, S., Hazratifard, S. M., & Dara, R. (2023). A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. IEEE Access.

152. Komaragiri, V. B., & Edward, A. (2022). AI-Driven Vulnerability Management and Automated Threat Mitigation. International Journal of Scientific Research.

153. Chinta, P. C. R., Jha, K. M., Velaga, V., & Moore, C. (2024). Harnessing Big Data and AI-Driven ERP Systems to Enhance Cybersecurity Resilience in Real-Time Threat Environments. SSRN Electronic Journal.

154. Humphreys, D., Koay, A., Desmond, D., & Mealy, E. (2024). AI Hype as a Cybersecurity Risk: The Moral Responsibility of Implementing Generative AI in Business. AI and Ethics, Springer.

Appendix B: Grey Literature References

1. OpenAI. (2023). Red Teaming Generative Models.
   https://openai.com/research/red-teaming-network
2. MITRE. (2023). ATLAS Matrix for AI Attacks.
   https://atlas.mitre.org/matrices/ATLAS
3. OWASP Foundation. (2023). Top 10 for LLM Applications.
   https://owasp.org/www-project-top-10-for-large-language-model-applications
4. Google. (2024). Secure AI Development Guidelines.
   https://ai.google/responsibility/security
5. IBM. (2024). Protecting AI Assets in Hybrid Clouds.
   https://www.ibm.com/security
6. CrowdStrike. (2024). Adversarial AI and Red Teaming Practices.
   https://www.crowdstrike.com
7. NVIDIA. (2023). Adversarial Defense Toolkit White Paper.
   https://developer.nvidia.com/adversarial-defense-toolkit
8. Microsoft. (2024). Responsible AI Security Testing Playbook.
   https://learn.microsoft.com/en-us/security/ai-security-playbook
9. Deloitte. (2024). AI in Critical Infrastructure Cybersecurity.
   https://www.deloitte.com
10. Accenture. (2024). AI Resilience in Healthcare and Finance.
    https://www.accenture.com
11. OECD. (2024). AI in Critical Systems: Safety and Governance.
    https://www.oecd.org
12. ENISA. (2024). AI in Energy and Transport Cybersecurity Report.
    https://www.enisa.europa.eu
13. Deloitte. (2024). Enterprise AI Security Blueprint.
    https://www.deloitte.com
14. IBM Security. (2024). AI for Threat Detection and Response.
    https://www.ibm.com/security
15. PwC. (2024). AI in Cybersecurity Operations Centers.
    https://www.pwc.com
16. Cisco. (2024). AI-Enhanced Threat Hunting.
    https://www.cisco.com
17. Capgemini. (2024). Securing the AI Enterprise.
    https://www.capgemini.com
18. NIST. (2023). AI Risk Management Framework (RMF).
    https://www.nist.gov/itl/ai-risk-management-framework
19. Cloud Security Alliance (CSA). (2024). AI Control Matrix.
    https://cloudsecurityalliance.org
20. OECD. (2024). Policy Framework for AI Security and Governance.
    https://www.oecd.org

21. Microsoft. (2024). AI Security Testing Toolkit. https://learn.microsoft.com/en-us/security

22. Zhang, T., Chen, J., & Wang, Y. (2023). Security and privacy risks in machine learning: A survey. https://www.semanticscholar.org/paper/SECURITY-AND-PRIVACY-IN-MACHINE-LEARNING%3A-A-SURVEY/fd031e917496b2566aa66e82d7730b39a87f3898

23. Paracha, A., Arshad, J., Ben Farah, M., & Ismail, K. (2024). Machine learning security and privacy: A review of threats and countermeasures, EURASIP Journal on Information Security. https://doi.org/10.1186/s13635-024-00158-3

24. Binns, R., et al. (2023). Data poisoning attacks and defenses in machine learning: A survey. https://arxiv.org/html/2503.22759v1

25. Koutnik, J., et al. (2023). Adversarial machine learning and defense mechanisms. https://www.springer.com/gp/book/978303108193

26. Kroll. (2023). AI security risks and recommendations. https://www.kroll.com/en/insights/publications/cyber/ai-security-risks-recommendations

27. Microsoft (2023). Six security considerations for machine learning solutions. https://techcommunity.microsoft.com/blog/fasttrackforazureblog/six-security-considerations-for-machine-learning-solutions/3718592

28. CyberProof. (2024). AI data security: Key threats and protection. https://www.cyberproof.com/blog/ai-data-security-key-threats-and-protection

29. IBM. (2023). Bringing light to the black box: AI governance and transparency. https://www.ibm.com/think/insights/bring-light-to-the-black-box

30. CyberProof. (2025). 2025 global threat intelligence report.

31. Accenture (2025). State of Cybersecurity 2025. https://www.accenture.com/us-en/insights/security/state-cybersecurity-2025

32. Schreiber, L., & Schreiber, M. (2025). AI for cyber-security risk: Harnessing AI for automatic generation of company-specific cybersecurity risk profiles, https://www.emerald.com/ics/article-abstract/33/4/520/1249885/AI-for-cyber-security-risk-harnessing-AI-for?redirectedFrom=fulltext&utm_source=researchgate

33. Lysenko, P., et al. (2024). The role of AI in cybersecurity: Automation of protection and detection, https://www.proquest.com/openview/ebe74758eaa95b3e9ea9e2882b9cfb1d/1?pq-origsite=gscholar&cbl=2032164

34. Rahman, A., et al. (2023). AI-powered solutions for enhancing national cybersecurity, https://www.researchgate.net/publication/387269707_AI-Powered_Solutions_for_Enhancing_National_Cybersecurity_Predictive_Analytics_and_Threat_Mitigation

35. Infosecurity Magazine. (2025). NIST warns of significant limitations in AI/ML security mitigations. https://www.infosecurity-magazine.com/news/nist-limitations-ai-ml-security/

36. Osler (2023), Emerging AI security risks and considerations: Key takeaways from the NIST adversarial machine learning report. https://www.osler.com/en/insights/updates/emerging-ai-security-risks-and-considerations-key-takeaways-from-the-nist-adversarial-machine-learning-report

37. Sarker, I. H. (2024). AI-driven cybersecurity and threat intelligence, https://link.springer.com/book/10.1007/978-3-031-54497-2

38. Jabbarova, N. (2023). AI and cybersecurity—New threats and opportunities, https://journlra.org/index.php/jra/article/view/754

39. OWASP(2025) Top 10 Risk & Mitigations for LLMs and Gen AI Apps, https://genai.owasp.org/llm-top-10/

40. Miessler, D. (2024). The AI attack surface map. https://danielmiessler.com/p/the-ai-attack-surface-map-v1-0

41. Billois, G., Bossuet, R., & Chardon, P.-L. (2024). Securing AI: The new cybersecurity challenges. https://www.riskinsight-wavestone.com/en/2024/03/securing-ai-the-new-cybersecurity-challenges

42. Business Insider. (2025). AI has ushered in a new kind of hacker. https://www.businessinsider.com/ai-hackers-models-hugging-face-opensource-jfrog-2025-3

43. Flynn, J., Rodriguez, A., & Popa, R. (2025). Evaluating potential cybersecurity threats of advanced AI. https://deepmind.google/discover/blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/

44. CrowdStrike. (2025). CrowdStrike Global Threat Report.

45. Deloitte. (2025). Annual cyber threat trends report. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-annual-cyber-threat-trends-report-2025.pdf

46. Cyber Threat Alliance. (2025). Cybersecurity in the age of generative AI. https://www.cyberthreatalliance.org/cyber-threat-alliance-publishes-2025-cybersecurity-in-the-age-of-generative-ai/

47. MIT Technology Review. (2025). The impact of AI on cybersecurity threats and defenses. https://www.technologyreview.com/2025/03/06/impact-of-ai-on-cybersecurity/

48. Rizvi, S. (2023). Enhancing cybersecurity: The power of artificial intelligence in threat detection and prevention, https://ijaers.com/uploads/issue_files/8IJAERS-05202313-Enhancing.pdf

49. Shah, P. (2021). Machine learning algorithms for cybersecurity, https://www.researchgate.net/publication/378396020_Machine_Learning_Algorithms_for_Cybersecurity_Detecting_and_Preventing_Threats

50. Radanliev, P., et al. (2024). AI security and cyber risk in IoT systems, https://arxiv.org/abs/2410.09194

51. Tetaly, P., & Kulkarni, S. (2022). Artificial intelligence in cyber security—A threat or a solution, https://pubs.aip.org/aip/acp/article-abstract/2519/1/030036/2828543/Artificial-intelligence-in-cyber-security-A-threat?redirectedFrom=fulltext

52. Labu, S., & Ahammed, M. (2024). Next-generation cyber threat detection and mitigation strategies, https://al-kindipublisher.com/index.php/jcsts/article/view/6803

53. DeepMind. (2024). Frontier AI risks: Misuse and security implications. https://deepmind.google/discover/blog/frontier-ai-risks

54. NIST (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations, https://csrc.nist.gov/pubs/ai/100/2/e2025/final

55. Hu, Y., et al. (2021). Artificial intelligence security: Threats and countermeasures. ACM Computing Surveys, 54(13s), Article 268,

https://www.researchgate.net/publication/366820580_Artificial_Intelligence_Security_Threats_and_Countermeasures

56. Elsevier Connect (2025). Rethinking peer review in the AI era: Responsibility and transparency. Elsevier Connect, https://www.elsevier.com/connect/rethinking-peer-review-in-the-ai-era-with-responsibility-and-transparency

57. CrowdStrike (2025). Modern Adversaries and Evasion Techniques: Why legacy antivirus is an easy target. CrowdStrike White Paper

58. Kawamoto, Y., et al. (2023) . Threats, Vulnerabilities & Controls of ML Systems, arX preprint, https://arxiv.org/abs/2301.07474

AI is transforming cybersecurity, and also reshaping its threat landscape. This report shows why securing AI demands more than patching, it requires resilience by design, transparency in deployment, and collaboration across sectors. The risks are real, the actors are diverse, and the tools are evolving. The time to act is now.

RISE – Research Institutes of Sweden
www.ri.se | info@ri.se | +46 10 516 50 00
Isafjordsgatan 22, SE-164 40 Kista
RISE Report: December 20, 2025