International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-15 Issue-4, November 2026

AI-Driven Cybersecurity in the Era of Autonomous Agents: Threat Detection, Automated Response, Adversarial Attacks, and AI Security

Jennifer Ikeoluwa Ade-omonijo, Micheal Abimbola Oladosu*, Moses Adondua Abah, Anya Adebayo Anya, Anyawuike Ikenna Omedobi, Emmanuel C. Ugwuabonyi

Abstract: ***Today, AI plays a pivotal role in cybersecurity, transforming how threats are identified, analysed, and mitigated. New autonomous, large language model (LLM)-based agents with the potential for independent reasoning and action create new defence capabilities and an expanded attack surface. Drawing on recent literature (2020-2025) across four interrelated themes, this review synthesises current knowledge on threat detection with artificial intelligence, automated incident response, adversarial attacks on AI-based security systems, and the security of AI and autonomous-agent systems. The models of machine learning and deep learning have found their way into use in intrusion detection, malware and phishing classification, and insider-threat analytics, and security orchestration, automation and response (SOAR) platforms are becoming more and more sophisticated to incorporate reinforcement learning and LLM-based agents to triage and remediate incidents with minimal human intervention. But as adversarial machine learning research has shown, these same models can be fooled, poisoned, and manipulated, and autonomous agents create new problems like 'misusing tools' and 'unsafe self-directed action'. As AI becomes increasingly sophisticated, the review finds an urgent need to develop robust, explainable, and continuously validated AI to close the gap between defences and attackers.***

Keywords: Artificial Intelligence, Cybersecurity, Autonomous Agents, Threat Detection, Adversarial Machine Learning, Automated Response.

Manuscript received on 12 October 2026 | First Revised Manuscript received on 17 October 2026 | Second Revised Manuscript received on 24 October 2026 | Accepted on 15 November 2026 | Manuscript published on 30 November 2026.

*Correspondence Author(s)
Jennifer Ikeoluwa Ade-omonijo, Department of Electrical and Information Engineering, Faculty of Engineering, Landmark University, Omu-Aran, Kwara State, Nigeria. Email ID: jenniferikeoluwa@gmail.com, ORCID ID: 0009-0000-9916-0749
Micheal Abimbola Oladosu*, Department of Chemical Sciences, Faculty of Sciences, Anchor University, Ayobo, Ipaja, Lagos, Nigeria. Email ID: mikeoladosu@gmail.com, moladosu@aul.edu.ng, ORCID ID: 0009-0000-5098-1247
Moses Adondua Abah, Department of Biochemistry, Faculty of Pure and Applied Sciences, Federal University of Wukari, Wukari, Taraba State, Nigeria. Email ID: m.abah@fuwukari.edu.ng, ORCID ID: 0000-0002-9268-1661
Anya Adebayo Anya, College of Management and Human Potential, Walden University, Minneapolis, United States. Email ID: anya.anya@walden.edu, ORCID ID: 0009-0006-9494-5313
Anyawuike Ikenna Omedobi, Department of Environmental Technology, Faculty of Engineering, Federal University of Technology Owerri, Imo State, Nigeria. Email ID: ikennaanyawuike@gmail.com, ORCID ID: 0009-0002-1378-0379
Emmanuel C. Ugwuabonyi, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, Maryland, United States. Email ID: ugwuabonyi@umbc.edu, ORCID ID: 0009-0006-3891-0001

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Nomenclature:
AI: Artificial Intelligence
DL: Deep Learning
LLMs: Large Language Models
SLR: Structured Literature Review
SOAR: Security Orchestration, Automation, and Response
NIDS: Network Intrusion Detection Systems
IoT: Internet of Things
AML: Adversarial Machine Learning

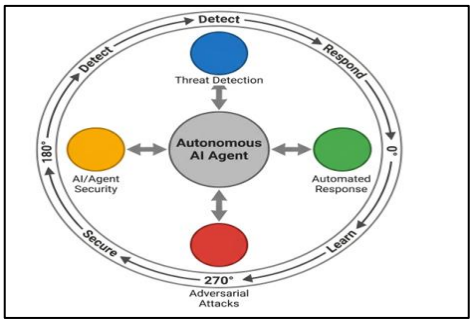
I. INTRODUCTION

The volume and sophistication of cyberattacks challenge traditional signature- and rule-based security systems, which struggle to keep pace with rapidly evolving threats. Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has become the dominant approach in modern security operations, enabling pattern detection in network traffic, endpoint behaviour, user activity, and more at a scale beyond human capability [1]. Beyond narrow, task-specific classifiers, the field is moving toward autonomous agents: systems that can plan, reason about, and execute multiple actions with limited human intervention, usually built on large language models (LLMs) [2].

This change is double-edged. Autonomous agents can shorten the time from detection to remediation and may automate security-orchestration tasks teams of analysts previously carried out around the clock [2,3]. However, the same generative and agentic abilities can be used in reverse: an LLM can generate phishing content, and agents can create new, poorly understood attack surfaces, such as prompt injection and tool misuse [3,4].

The four strands of research that need to be explored when understanding this landscape are: (i) how AI can enhance the detection of malicious activity; (ii) how AI allows for faster and more autonomously driven incident response; (iii) how adversaries can exploit weaknesses in AI-based defences through adversarial machine learning; and (iv) how the security of AI-based and autonomous-agent systems can be evaluated and protected. This review synthesises the existing literature on AI in cybersecurity from the past five years (2020–2025), published in peer-reviewed journals and conference proceedings, into a single, state-of-the-art, evidence-based report. Figure 1 presents the conceptual taxonomy of autonomous agents in AI-driven cybersecurity.

© Copyright: All rights reserved.



[Fig.1: Conceptual Taxonomy of AI-Driven Cybersecurity in the Era of Autonomous Agents]

Source: Authors' Own Conceptual Synthesis, Based on the Four-Domain Structure of this Review.

II. METHODOLOGY

A. Review Design

This study used a Structured Literature Review (SLR) to synthesise the latest research on artificial intelligence (AI)-based cybersecurity in the autonomous agent era. These four areas were selected for the review due to their interrelatedness: (i) AI-based cyber threat detection, (ii) automated incident response and autonomous cyber-defence, (iii) adversarial attacks on AI-based cybersecurity systems, and (iv) security, trustworthiness, and explainability of AI and autonomous-agent systems. The review aimed to provide a holistic view of current capabilities, limitations, vulnerabilities, and promising future research areas.

B. Scope and Eligibility of the Literature

The review examined peer-reviewed literature published between 2020 and 2025. The chosen time frame was designed to reflect the swift pace of innovation in cybersecurity, particularly with the use of machine learning and deep learning technologies, as well as newer advancements in the use of large language models (LLMs), generative AI, reinforcement learning, security orchestration and automation and response (SOAR), and autonomous AI agents.

We deemed studies that discussed any of the four main themes of the review relevant. For AI-based threat detection, we focused on network intrusion detection, malware detection, phishing detection, insider-threat detection, and AI-based detection in IoT and industrial IoT environments. The studies analysed for automated responses included SOAR, reinforcement-learning-based cyber defence, autonomous security agents, and LLM-based security operations. Under adversarial machine learning, we included literature covering evasion, poisoning, model extraction, adversarial attacks, and countermeasures. The AI security domain included studies on security, explainability, robustness, governance, and the safe deployment of AI and autonomous agents.

C. Literature Selection and Thematic Organisation

The review grouped the literature into themes based on its four-domain conceptual framework. The review examined relevant studies and identified common trends, technological advances, reported benefits, limitations, security issues, and unresolved research questions.

The review emphasised studies that provide evidence for the effectiveness and limitations of AI-based cybersecurity methods. For example, it reported improvements in detection accuracy, reduced false positives, faster response times, and AI systems' ability to detect new threats. It also considered challenges such as reliance on outdated benchmark datasets, computational demands, class imbalance, privacy concerns, susceptibility to adversarial attacks, model drift, and deployment challenges in real-world environments.

D. Analytical Framework

The study compared findings from the four main research areas using thematic synthesis. In the context of AI-based threat detection, the analysis focused on machine learning, deep learning, ensemble methods, recurrent and convolutional architectures, transformer-based approaches, and federated or distributed learning.

The analysis explored automated response and autonomous defence through AI-enhanced SOAR, reinforcement learning defence agents, causal reinforcement learning, and autonomous agents using LLMs. They analysed these approaches for their roles in alert triage, threat prioritisation, containment, remediation, and multi-step security operations.

For adversarial attacks, studies were grouped by attack mechanism and the targeted phase of the machine-learning lifecycle. These encompassed evasion attacks, poisoning attacks, model extraction/inversion attacks, and attacks targeting IoT-based AI security systems. The analysis also covered corresponding defensive methods such as adversarial training, input sanitisation, ensemble methods, anomaly detection, and query monitoring.

The analysis covered prompt injection, insecure output handling, training-data poisoning, adversarial manipulation, tool misuse, explainability, constrained tool permissions, human oversight, and auditability for AI and autonomous-agent security. This allowed the review to evaluate not only autonomous AI's potential in cybersecurity, but also the dangers of entrusting AI systems with greater autonomy at each stage.

E. Evidence Synthesis

The selected literature was not statistically pooled but analysed qualitatively. The synthesis examined the four research domains and included reported approaches, applications, benefits, limitations, and security implications. The synthesis consolidated a representative set of AI/ML techniques, adversarial attack categories, and automated response approaches into tables and created conceptual figures showing relationships among threat detection, automated response, adversarial attacks, and AI-agent security.

The synthesis also explored the disconnect between controlled or benchmark environments and real-world cybersecurity needs. It focused on robustness, explainability, continuous validation, human oversight, and the safe deployment of autonomous agents.

F. Quality and Relevance Considerations

The review prioritised peer-reviewed research directly relevant to the study themes and offering substantive evidence on AI-driven cybersecurity technologies, their effectiveness, vulnerabilities, or defences. The review paid special attention to systematic reviews, comprehensive

© Copyright: All rights reserved.

survey research, empirical studies, and research that discussed or proposed a specific AI/ML architecture or cybersecurity framework.

The review also considered limitations frequently noted in the literature, such as benchmark dependence, limited generalisability to production settings, computational limitations, adversarial vulnerability, privacy concerns, and the lack of effective governance for autonomous AI systems.

G. Conceptual Synthesis

The final synthesis consolidated the four domains' outputs into a comprehensive concept for an AI-driven cybersecurity approach. The model treats modern AI in cybersecurity as a dynamic, ongoing process characterised by threat detection, automated response, adversarial adaptation, and safeguarding AI systems. This viewpoint acknowledges not only the need for ongoing validation, adversarial robustness, explainability, constrained autonomy, and human oversight, but also that increased defensive automation introduces new attack surfaces.

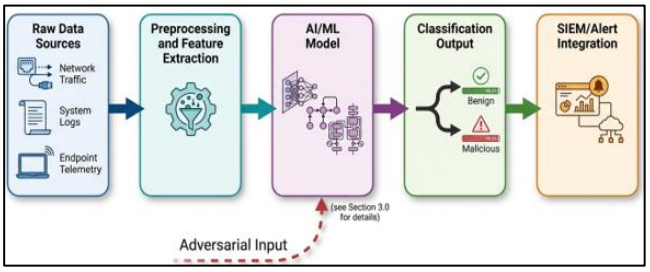
III. AI-DRIVEN THREAT DETECTION

Current research trends in network intrusion detection systems (NIDS) focus on machine learning and deep learning methods. These systematic reviews show that convolutional and recurrent architectures outperform classical algorithms on benchmark datasets in every case. However, performance improvements are often reported only in lab settings, on older datasets that do not reflect production network traffic [1,5,6]. Ensemble and hybrid architectures further improve detection accuracy and reduce false positives by adding feature-selection steps before applying deep classifiers, which is problematic in resource-constrained Internet of Things (IoT) deployments [5]. Trust and robustness in distributed/federated network settings continue to be issues of interest, especially when models are trained collaboratively among nodes that are not trusted [7].

The same trend has been seen with malware detection. This is because deep learning classifiers, particularly convolutional networks that use binary visualisation and recurrent models that use API call traces, have significantly improved detection of previously unseen (zero-day) malware families compared to signature-based AV engines [8]. Comprehensive reviews, however, indicate that adversarial evasion (Section 3.0) and ever-diversifying malware families, whether in desktop, mobile, or IoT contexts, remain ahead of the curve and warrant a shift toward regularly updated ensemble and transfer-learning-type detectors [9].

For phishing and social-engineering detection, the shift from blacklist-based heuristics to machine-learning classifiers trained on URL structure, page content, and behavioural features has been equally successful, with high accuracy. Still, classifiers remain vulnerable to fast-rotating infrastructures and increasingly sophisticated, AI-generated lure content [10]. Advances in anomaly detection and sequence modelling have improved insider-threat detection, which relies on behavioural and access-log analytics rather than network signatures; however, reviewers continually identify class imbalance, base-rate issues, and privacy concerns about monitoring worker activity as challenges [11].

Another recent trend is using LLMs directly in threat detection. Researchers have adapted and optimised transformer-based language models for network traffic, highlighting their potential for structured security telemetry while addressing edge-deployment constraints such as latency and the limited footprint of lightweight devices, including those in the IoT and industrial IoT (IIoT) sectors [12]. Table 1 summarises representative AI/ML techniques applied across these threat-detection domains, their principal strengths, and their most commonly reported limitations. Figure 2 presents the Generalised AI/ML Pipeline for Cyber Threat Detection.



[Fig.2: Generalised AI/ML Pipeline for Cyber Threat Detection]

Source: Authors' Own Synthesis of the Detection Pipeline Described in References [1,6].

Table I: AI/ML Techniques for Cyber Threat Detection

Detection Domain	Representative AI/ML Technique	Reported Strength	Key Limitation	Ref.
Network intrusion detection	CNN/RNN and hybrid ensembles	High accuracy on benchmark datasets; strong pattern recognition	Reliance on ageing datasets; limited generalisation to live traffic	[1,5,6]
IoT/SDN-based intrusion detection	SDN-integrated deep learning	Centralised visibility across distributed IoT nodes	High computational demand for constrained devices	[5]
Distributed/federated networks	Federated trust-aware AI	Preserves data locality; scalable trust scoring	Vulnerable to poisoning by malicious participants	[7]
Malware detection	CNN on binary visualisation; RNN on API sequences	Detects previously unseen (zero-day) variants	Susceptible to adversarial evasion; concept drift	[8,9]
Phishing detection	URL/content feature classifiers (SVM, RF, DL)	High accuracy on labelled corpora	Rapid rotation of infrastructure; AI-generated lures	[10]
Insider-threat detection	Behavioural sequence/anomaly models	Captures subtle deviations in user behaviour	Severe class imbalance; monitoring/privacy constraints	[11]
Lightweight IoT/IIoT threat detection	Fine-tuned transformer (BERT-based)	Low latency, small footprint suitable for edge devices	Requires domain-specific fine-tuning and retraining	[12]

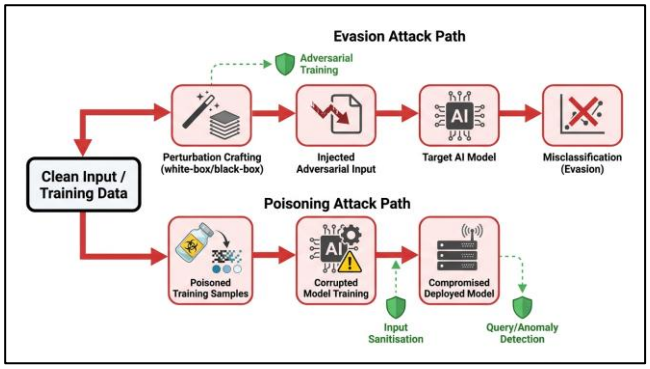
© Copyright: All rights reserved.

IV. ADVERSARIAL ATTACKS AGAINST AI-BASED SECURITY SYSTEMS

The same statistical learning properties that make ML and DL models effective detectors also make them exploitable. Adversarial machine learning (AML) research identifies various stages of the machine learning lifecycle that can be attacked, the kind of prior information the attacker has about the model (e.g., white box or black box), and the type of attack that can be conducted (evasion, poisoning, or model extraction) [13,14]. Small perturbations can be added to malicious traffic input to evade detection by a deep learning-based NIDS; for example, it is demonstrated that adversarial training, ensemble voting or query detection can successfully increase the evasion rate of near-complete adversarial training for the deep learning-based NIDS against small perturbations of malicious traffic input, unless they are applied [15,16]. This sub-field note reflects the complexity of building adversarial examples in the network domain: while it is more challenging than in the image domain, it is not impossible, and successful attacks against widely used, benchmark-trainable detectors have been demonstrated [16,17].

This risk is exacerbated in the Internet of Things, where resource-constrained IoT security systems most likely rely on lightweight ML classifiers and least likely have the computational resources to implement expensive adversarial-defence methods like adversarial training or ensembling [18]. Systematic reviews of AML in IoT-enabled security systems identify many attack vectors applicable to intrusion detection, device fingerprinting, and access control models for IoT systems, but relatively few lightweight, IoT-friendly defences [19].

In addition to intrusion detection, several other defensive controls have been documented in AML literature within the broader cyber security field, such as adversarial training, input sanitisation, gradient masking, certified robustness bounds and anomaly-based detection of adversarial queries, which are all mechanisms to prevent intrusion, but which are limited in their ability to generalise across different types of attacks and domains [20]. The collective evidence suggests treating adversarial robustness on the same level as other design factors for an AI-powered security system, not as an add-on feature to an existing detector. Table 2 summarises the main categories of the most common adversarial attacks reported in the literature, their mechanism and representative countermeasures. Figure 3 presents the adversarial attack lifecycle against AI-based security models.



[Fig.3: Adversarial Attack Lifecycle Against AI-Based Security Models]

Source: Authors' Own Synthesis, Based on the Attack Taxonomies of References [13,15].

Table II: Categories of Adversarial Attacks Against AI-Based Security Systems

Attack Category	ML Lifecycle Stage	Mechanism	Representative Defence	Ref.
Evasion attacks	Inference/deployment	Perturb malicious input to avoid classification	Adversarial training; ensemble voting	[13,15,16]
Poisoning attacks	Training	Inject manipulated samples into training data	Data sanitisation; robust/anomaly-aware training	[13,14]
Model extraction/inversion	Post-deployment (query-based)	Reconstruct model or training data via repeated queries	Query-rate limiting; output perturbation	[14,20]
IoT-targeted adversarial attacks	Inference (resource-constrained)	Lightweight perturbations exploiting shallow IoT classifiers	Lightweight/quantised adversarial defences	[18,19]
Domain-constrained network evasion	Inference	Statistically valid perturbations preserving traffic semantics	Feature-space constraints; query detection	[15,16,17]

V. AUTOMATED RESPONSE AND AUTONOMOUS CYBER-DEFENCE AGENTS

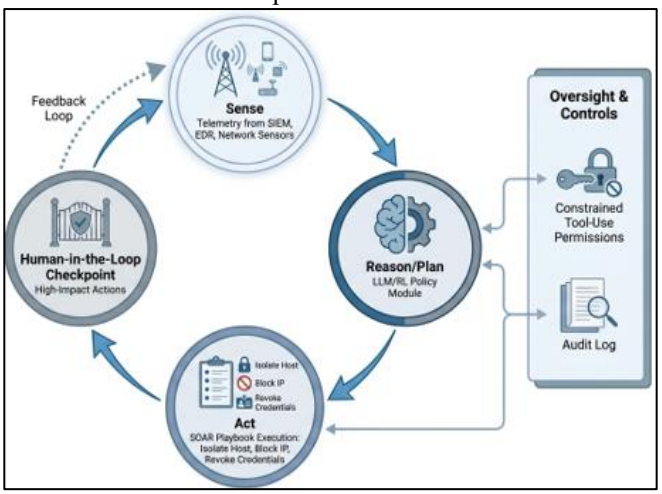
The volume of detections has surpassed analysts' capacity, and security orchestration, automation, and response (SOAR) platforms are leveraging AI/ML to prioritise alerts, suggest playbooks, and increasingly perform containment actions automatically. Alert triage, playbook recommendation, and false-positive suppression are the most developed AI/ML use cases in SOAR [21]. At the same time, operational risk limits fully automatic remediation in production [5] and [12]. Novel SOAR architectures for new, increasingly heterogeneous (IT/OT/IoT) environments expand this orchestration logic beyond previously siloed security domains, unifying response logic across cloud, on-premises, and edge assets in a single automation layer [22].

A recent and active research area for training autonomous cyber-defence agents is reinforcement learning (RL), which learns optimal response policies through simulated interaction with attackers, without being programmed with a response playbook [23]. RL-trained

© Copyright: All rights reserved.

agents can reliably defend network testbeds against network-wide, advanced persistent threat-style adversaries, such as those organised in a hierarchical architecture that incorporates specialised sub-agents controlled by a higher-level cyber-defence controller [24]. More recent attempts add causal reasoning to the RL reward and state representation, helping agents generalise defensive policies to previously unknown attack policies and provide more interpretable justifications for their actions [25].

Most recently, LLM-based autonomous agents that support multi-step planning, tool use, and natural-language reasoning about a security context have begun to be applied throughout the offence-defence lifecycle, from automated penetration testing to incident triage and playbook generation [2]. Initial assessments indicate that these agents could have a significant impact on the mean time to detect and mean time to respond when incorporated into the security-operations-centre workflow, but that they also depend on the reliability of models, the risk of hallucination, and the requirement for human-in-the-loop verification before the implementation of high-impact actions [4]. Table 3 shows representative examples of automated-response and autonomous-agent approaches, the techniques used, and each approach's main purpose in the incident-response lifecycle. Figure 4: Architecture of an Autonomous AI Agent for Automated Threat Detection and Response:



[Fig.4: Architecture of an Autonomous AI Agent for Automated Threat Detection and Response]

Source: Authors' Own Synthesis, based on the Agent Frameworks Described in References [2,21].

Table III: Automated Response and Autonomous Agent Approaches in Cybersecurity

Approach	Underlying Technique	Primary Function	Reported Benefit	Ref.
AI/ML-enhanced SOAR	Supervised ML alert scoring and playbook recommendation	Alert triage and prioritisation	Reduced analyst workload; faster mean-time-to-triage	[21]
Blended-environment SOAR	Cross-domain orchestration architecture	Coordinated response across IT/OT/IoT assets	Unified automation across previously siloed domains	[22]
Reinforcement-learning defence agents	Deep/hierarchical RL	Learn optimal containment/response policy via simulation	Reliable defence against simulated APT-style adversaries	[23,24]
Causally-aware RL agents	Causal reinforcement learning	Generalisable, interpretable defensive policy selection	Improved generalisation to unseen attack sequences	[25]
LLM-based autonomous security agents	LLM with tool use/planning	Multi-step reasoning across detect-to-remediate lifecycle	Reduced mean time to detect/respond; automated triage	[2,4]

VI. AI SECURITY IN THE ERA OF AUTONOMOUS AGENTS

As AI security tools become more autonomous, the stakes have risen: the confidentiality, integrity, and trustworthiness of these models and agents increasingly shape consequential defensive (and possibly offensive) decisions. In-depth analyses of generative AI and LLMs in cybersecurity reveal a "double use" scenario in which the same AI systems used for improved threat intelligence, malware analysis, and secure code review can be weaponised via prompt injection, improper output handling, training data poisoning, and adversarial prompts to generate malicious or dangerous content [3]. Unlike a standalone chatbot, LLM output from a prompt injection or jailbreak can be chained into further actions that lead to unauthorised system changes, representing a qualitative risk shift compared with a standalone chatbot [4].

Explanatory AI (XAI) can partially address opacity. Research on XAI in cybersecurity shows a growing use of model-agnostic approaches to explain AI decisions, such as feature-attribution and rule-extraction methods, alongside a new line of research on attacks on the explanation pipeline itself [26,27]. For autonomous agents, the explainer's explainability matters, because an agent's internal reasoning traces may be its only explanation for the action it took.

Collectively, the literature suggests that securing AI in the agentic era requires layered assurance: adversarially robust base models (Section 3.0), constrained and auditable tool-use permissions for autonomous agents, human-in-the-loop checkpoints for high-impact actions, and explainability mechanisms that are themselves protected from manipulation. The reviewed literature also points to governance frameworks and technical measures meant to discipline these systems, developed significantly slower than agent deployment, leaving the major open problem in AI-powered cybersecurity as governing what is being deployed.

VII. CONCLUSION

From supportive analytics tool to its current role in today's cybersecurity

© Copyright: All rights reserved.

environment, AI's shift to autonomous decision-making is the next major evolution. While it is true that the accuracy of detection and the speediness of response are measurable, it is also true that there remains a measurable difference between the sophistication of the defence and the sophistication of the adversary: the same architectures that detect threats can be evaded, poisoned, or manipulated, and the agents built to respond autonomously introduce new categories of risk into which traditional security engineering has yet to fit fully. This requires treating adversarial robustness, explainability, and constrained autonomy as core design considerations, and empirically testing autonomous agents in both realistic and adversarial settings beyond benchmarks. The question is no longer whether AI can detect and respond to threats, but whether it can do so safely, transparently and robustly at scale.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of the Authors who assisted with search strategy development and all researchers whose work was included in this review.

DECLARATION STATEMENT

After aggregating input from all authors, I, as the article's author, must verify the accuracy of the following information.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures the research is conducted objectively, without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The data and materials in this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCE

1. Ahmad Z, Khan AS, Shiang CW, Abdullah J, Ahmad F. Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Trans Emerg Telecommun Technol. 2021;32(1): e4150. DOI: 10.1002 /ett.4150
2. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model-based autonomous agents. Front Comput Sci. 2024;18(6):186345. DOI: 10.1007/s11704-024-40231-1
3. Ferrag MA, Alwahedi F, Battah A, Cherif B, Mechri A, Tihanyi N, et al. Generative AI in cybersecurity: a comprehensive review of LLM applications and vulnerabilities. Internet Things Cyber-Phys Syst. 2025; 5:1-46. DOI:10.1016/j.iotcps.2025.01.001
4. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access. 2023; 11:80218-45. DOI:10.1109/ACCESS.2023.3300381
5. Muthanna MSA, Alkanhel R, Muthanna A, Rafiq A, Abdullah WAM. Towards an SDN-enabled, intelligent intrusion detection system for the Internet of Things (IoT). IEEE Access. 2022; 10:22756-68. DOI: 10.1109/ACCESS.2022.3153716
6. Momand A, Jan SU, Ramzan N. A systematic and comprehensive survey of recent advances in intrusion detection systems using machine learning: deep learning, datasets, and attack taxonomy. J Sens. 2023; 2023:6048087. DOI: 10.1155/2023/6048087
7. Li Z, Fang W, Zhu C, Gao Z, Zhang W. AI-enabled trust in distributed networks. IEEE Access. 2023; 11:88116-34. DOI: 10.1109/ACCESS.2023.3306452
8. Bensaoud A, Kalita J, Bensaoud M. A survey of malware detection using deep learning. Mach Learn Appl. 2024; 16:100546. DOI: 10.1016/j.mlwa.2024.100546
9. Maniriho P, Mahmood AN, Chowdhury MJM. A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms. ACM Comput Surv. 2024;56(6):145. DOI: 10.1145/3638240
10. Tang L, Mahmoud QH. A survey of machine learning-based solutions for phishing website detection. Mach Learn Knowl Extr. 2021;3(3):672-94. DOI: 10.3390/make3030034
11. Alzaabi FR, Mehmood A. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. IEEE Access. 2024; 12:30907-27. DOI: 10.1109/ACCESS.2024.3369906
12. Ferrag MA, Ndhlovu M, Tihanyi N, Cordeiro LC, Debbah M, Lestable T, et al. Revolutionising cyber threat detection with large language models: a privacy-preserving BERT-based lightweight model for IoT/IIoT devices. IEEE Access. 2024; 12:23733-50. DOI: 10.1109/ACCESS.2024.3363469
13. Rosenberg I, Shabtai A, Elovici Y, Rokach L. Adversarial machine learning attacks and defence methods in the cyber security domain. ACM Comput Surv. 2021;54(5):108. DOI: 10.1145/3453158
14. Alotaibi A, Rassam MA. Adversarial machine learning attacks against intrusion detection systems: a survey on strategies and defence. Future Internet. 2023;15(2):62. DOI: 10.3390/fi15020062
15. Zhang C, Costa-Pérez X, Patras P. Adversarial attacks against deep learning-based network intrusion detection systems and defence mechanisms. IEEE/ACM Trans Netw. 2022;30(3):1294-311. DOI: 10.1109/TNET.2021.3137084
16. He K, Kim DD, Asghar MR. Adversarial machine learning for network intrusion detection systems: a comprehensive survey. IEEE Commun Surv Tutor. 2023;25(1):538-66. DOI: 10.1109/COMST.2022.3233793
17. Alshahrani E, Alghazzawi D, Alotaibi R, Rabie O, Alzubi OA. Adversarial attacks against supervised machine learning-based network intrusion detection systems. PLoS One. 2022;17(10): e0275971. DOI:10.1371/journal.pone.0275971
18. Khazane H, Ridouani M, Salahdine F, Kaabouch N. A holistic review of machine learning adversarial attacks in IoT networks. Future Internet. 2024;16(1):32. DOI: 10.3390/fi16010032
19. Harbi Y, Medani K, Gherbi C, Aliouat Z, Harous S. Roadmap of adversarial machine learning in Internet of Things-enabled security systems. Sensors. 2024;24(16):5150. DOI: 10.3390/s24165150
20. Malik J, Muthalagu R, Pawar PM. A systematic review of adversarial machine learning attacks, defensive controls, and technologies. IEEE Access. 2024; 12:99382-421. DOI: 10.1109/ACCESS.2024.3423323
21. Kinyua J, Awuah L. AI/ML in security orchestration, automation and response: future research directions. Intell Autom Soft Comput. 2021;28(2):527-45. DOI: 10.32604/iasc.2021.016240
22. Lee M, Jang-Jaccard J, Kwak J. Novel architecture of security orchestration, automation and response in internet of blended environment. Comput Mater Contin. 2022;73(1):199-223. DOI: 10.32604/cmc.2022.028495
23. Adawadkar AMK, Kulkarni N. Cyber-security and reinforcement learning—a brief survey. Eng Appl Artif Intell. 2022; 114:105116. DOI:10.1016/j.engappai.2022.105116
24. Foley M, Hicks C, Highnam K, Mavroudis V. Autonomous network defence using reinforcement learning. In: Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22). New York: Association for Computing Machinery; 2022. DOI: 10.1145/3488932.3527286
25. Purves T, Kyriakopoulos KG, Jenkins S, Phillips I, Dudman T. Causally aware reinforcement learning agents for autonomous cyber defence. Knowl Based Syst. 2024; 304:112521. DOI: 10.1016/j.knosys.2024.112521
26. Charmet F, Tanuwidjaja HC, Ayoubi S, Gimenez PF, Han Y, Jmila H, et al. Explainable artificial intelligence for cybersecurity: a literature survey. Ann Telecommun. 2022;77(11-12):789-812. DOI: 10.1007/s12243-022-00926-7
27. Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in

© Copyright: All rights reserved.

cybersecurity: a survey. IEEE Access. 2022; 10:93575-600. DOI: 10.1109/ACCESS.2022.3204171

AUTHOR'S PROFILE



Jennifer Ikeoluwa Ade-omonijo is an Electrical Engineering graduate and researcher with a strong interest in artificial intelligence, intelligent systems, human-computer interaction, and digital technologies. She holds a Bachelor's degree in Electrical and Information Engineering from Landmark University, Nigeria. She has gained experience through academic research, engineering practice, and professional training in wireless communication, machine learning, digital systems, and user-centred design. Her research interests include artificial intelligence, digital twins, human-AI interaction, intelligent systems, wireless communication, and applying emerging technologies to real-world challenges. She is also developing expertise in UI/UX and Product Design, with a particular interest in creating accessible, user-centred, and innovative digital experiences. She is passionate about interdisciplinary research and contributing to technological solutions that improve how people interact with digital and intelligent systems.



Micheal Abimbola Oladosu is an interdisciplinary researcher with a strong academic and research background. He founded Research Hub Nexus Institute, a platform dedicated to research publication, academic writing support, and mentorship for emerging scholars. He currently lectures in the Department of Chemical Sciences (Biochemistry Unit), Anchor University, Lagos, Nigeria. His research interests span the intersections of the life sciences, biomedical sciences, and technology, and he has actively collaborated on multidisciplinary research projects published in reputable peer-reviewed journals. He also reviews for several academic journals and serves on the editorial boards of multiple international scientific journals and publications.



Moses Abah Adondua is a dedicated biochemist and lecturer at the Department of Biochemistry, Faculty of Biosciences, Federal University Wukari, Nigeria. With a strong academic foundation and research interests in molecular biology, enzymology, and metabolic disorders, he teaches, conducts research, mentors undergraduate students, and helps develop innovative approaches to biochemistry research. research in biochemistry. He is currently the Co-Founder of Research Hub Nexus, where he helps advance research, innovation, and academic collaboration. He also participates in scholarly publishing, peer review, and capacity-building initiatives aimed at strengthening scientific research in Africa and globally, respectively.



Anya Adebayo is a Political Science graduate from Obafemi Awolowo University, Ile-Ife, Nigeria, with interdisciplinary research interests spanning artificial intelligence, robotics, public policy, security studies, sustainable development, and technological innovation. His academic and professional experience spans development research, governance, strategic analysis, emerging technologies, and policy-oriented research, with a particular emphasis on applying Artificial Intelligence, Machine Learning, Robotics, Blockchain, and Internet of Things (IoT) technologies to contemporary development and security challenges in Africa. His research spans diverse interdisciplinary areas, including climate resilience, public health surveillance, sustainable agriculture, waste management, urban security, disaster resilience, electoral integrity, and development policy. His scholarly work examines how emerging technologies and evidence-based policy frameworks can strengthen institutional capacity, improve societal resilience, and support sustainable development outcomes. His current research interests include AI-enabled governance and security, intelligent and autonomous systems, robotics, climate and disaster resilience, public health intelligence, sustainable development, technology policy, and the intersection of emerging technologies with development and security studies. His broader scholarly focus is on advancing interdisciplinary approaches that leverage technological innovation, policy research, and data-driven methodologies to address complex socioeconomic and developmental challenges in Africa.



Anyawuike Ikenna Omedobi holds a B. Tech in Geo-Environmental Technology from the Federal University of Technology Owerri (FUTO), Nigeria. He is a Data Scientist and Machine Learning Engineer with experience spanning data science, machine learning, environmental technology, and financial analytics. He currently works with Note.AI Ltd, Hull, UK, where he applies data-driven approaches to financial analysis, credit risk, and business decision-making. His technical expertise includes Python, SQL, exploratory data analysis, predictive modelling, machine learning, deep learning, computer vision, and MLOps. His research interests include environmental and financial applications of artificial intelligence, predictive modelling, computer vision, natural language processing, and deep learning for cybersecurity, critical infrastructure and financial security systems. He holds professional certifications in Data Science from IBM, Tensor Developer from DeepLearning.ai, and has completed AI and data analytics programmes with BCG X, KPMG, and other leading organisations.



Emmanuel C. Ugwuabonyi is a PhD researcher at the University of Maryland, Baltimore County, where he applies artificial intelligence and causal inference to healthcare challenges. His work spans machine learning for clinical and physiological data, the reliability, robustness, and safety of AI systems deployed in medical settings, and algorithmic fairness in health-related decision-making. His academic and professional experience spans Nigeria, Germany, and the United States, giving him an international perspective on developing and responsibly applying data-driven technologies in healthcare.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Publisher /Journal and/or the editor(s). The Publisher /Journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

© Copyright: All rights reserved.