ETDT

Emerging Trends in Digital Transformation

Volume-1, Issue-3, November-2025 ISSN: **3107-4308** *Paper ID:* ***ETDT-V1-I3-04***

AI-Driven Cybersecurity Frameworks for Intelligent Detection, Defense, and Response to Autonomous Threats

[1]George Zacharia

[1]Independent Researcher Mechanicsburg, PA, USA

E-mail: georgezacharia1983@gmail.com.

Abstract

Artificial intelligence (AI) now has the power to change the face of cybersecurity, shifting from rule-based, reactive protection to adaptive, autonomous, and intelligence-driven defence. This review highlights the impact of AI on cybersecurity and the role of intelligent autonomous systems in threat detection, adaptive defense, and autonomous incident response, with a special focus on agentic AI systems that could be tasked with, or required to, reason, learn and act under constantly evolving threat conditions. The review provides a snapshot of the latest advancements in machine learning and deep learning, behavioural analytics, threat intelligence, automated orchestration, and autonomous security agents, along with the increasing pace of AI-driven attacks and the speed of attacks. A special focus is placed on the shift from conventional security automation to cognitive and self-adaptive security defence strategies, which can detect new or abnormal behaviour, prioritise threats, and trigger containment or remediation. The analysis also reveals some of the most important drawbacks of adversarial manipulation, model opacity, false positives, reliable autonomous decisions, accountability, and the lack of common evaluation metrics. Several requirements are highlighted to ensure deployment in a "trustworthy" manner, including human-in-the-loop oversight, explainable AI, adversarial robustness, formal verification, and adaptive governance. A layered conceptual framework for combining intelligent detection, contextual risk assessment, adaptive defense, Autonomous response, continuous learning and good governance is offered based on the above evidence. A balance among machine speed, metastability, explainability, auditability, human involvement, and regulation will be a key feature of future cybersecurity autonomics systems to guarantee secure and resilient cyber defence, concludes the review.

Keywords: Agentic AI, Artificial Intelligence, Autonomous Cyber Defense, Cybersecurity, Intelligent Threat Detection, Adaptive Defense, Autonomous Response, Explainable AI.

Review Strategy and Evaluation Criteria

This review explores the evolution of the use of autonomous agentic systems in cyber protection, from reactive automated approaches to intelligent, cognitive reasoning structures (ADABARA et al., 2025). This includes assessing how AI agents can help combat new and complex AI-driven threats that evade traditional, passive security systems (Bhargava & Bommela, 2025; Malatji & Tolah, 2024). To enable a thorough examination, this review adopts a multifaceted analytical framework that focuses on operational efficacy, adaptive learning capabilities, and the intrinsic trustworthiness of autonomous decision-making systems (Erigha et al., 2025; Malatji, 2025). The criteria also need to be negotiated in light of the requirement for human oversight and control, as part of preserving machine autonomy and ensuring regulatory and ethical management (Al-E'mari et al., 2025; Dehghantanha et al.,

DOI: https://doi.org/10.65713/etdtv1i304

2023). This research aims to dissect the reasons behind these essential performance metrics and understand how AI can effectively support predicting and countering complex, machine-speed attacks (Shahana et al., 2024). The assessment is focused on the systemic analysis of machine learning and metaheuristic techniques with their ability to willingly preserve integrity in the face of adversarial AI (AI) growing at a fast pace (Salem et al., 2024).

Conceptual Foundations

In the context of modern operations, the use of artificial intelligence, particularly AI models like machine learning and deep learning, could help in proactively identifying new threats, including zero-day exploits and advanced persistent threats (APTs) (Lurds, 2026). These intelligent systems are designed to detect anomalies in real time, which is a key advantage over traditional signature-based approaches.One major advantage of these intelligent systems over traditional signature-based approaches is their ability to detect anomalies in real time. Moreover, they stand out from fixed ones because they adaptively protect against and recover from threats at machine speed, relying on adaptive control loops and self-healing capabilities to bolster system resilience (Al-E'mari et al., 2025). What is important to recognize is that autonomous threats work differently from the traditional model of threats—by dynamically changing its courses of attack, such an approach requires us to shift to using agentic AI that can actively simulate and offset such behaviors (Panwar & Abdelrahman, 2025; Sabillón et al., 2024).

Review Framework

Security architectures are classified by their intelligence, ranging from playbook-based automation to reasoning under uncertainty in advanced agentic security systems (Devagiri, 2026a, 2026b). Operating functions are then broken down into their ability to perform closed-loop lifecycle—including detection, confirmation, remediation and validation—to combat the "silo mentality" of stand-alone defense instruments (Khurana & Jain, 2025). We use performance metrics of system adaptability, computational latency, and decision making in the face of adversarial conditions to assess these frameworks (Devagiri, 2026; Kukkadapu, 2025).

Table 1. Evolution of AI-Driven Cybersecurity Frameworks

Framework Level	Core Approach	Detection Capability	Defense/Response	Adaptability	Human Role
Rule-Based Security	Predefined rules and signatures	Known threats	Manual or predefined actions	Very low	High
Automated Security	Playbooks and workflow automation	Known and predefined anomalies	Automated routine actions	Low	High
ML-Driven Security	Supervised/unsupervised learning	Known and emerging patterns	Risk-based automated actions	Moderate	Moderate
Adaptive AI Defense	Continuous analytics and contextual learning	Dynamic and evolving threats	Adaptive policy enforcement	High	Supervisory

DOI: https://doi.org/10.65713/etdtv1i304

Framework Level	Core Approach	Detection Capability	Defense/Response	Adaptability	Human Role
Agentic AI Cyber Defense	Autonomous reasoning, planning, and learning	Complex and autonomous threats	Autonomous containment and remediation	Very high	Governance and high-risk approval

Autonomous Threat Landscape

Today's attack surface is characterised by autonomous actors who initiate, execute, and adapt their attacks without human involvement (Schmitt & Koutroumpis, 2025). Such in-game strategies include leveraging advanced planning, memory augmentation, and feedback mechanisms to outsmart passive protection—an expression of autonomy used as a tool to achieve intricate goals (ADABARA et al., 2025; Deng et al., 2025). These entities enable autonomous exploration and tool use, going beyond simple one-step executions, reasoning over a period of deliberation, exploitation, and identification through perceptual, cognitive, and executive layers (Lazer et al., 2026; Radanliev et al., 2026). These are enemies that incorporate descending cycles of decision-making and utilise tools to perform quick reconnaissance and adapt to advance, constantly evolving how they manoeuvre to avoid discovery (Lazer et al., 2026; Li & Zhu, 2025).

Table 2. Major Autonomous and AI-Enabled Cyber Threats and Corresponding Defense Requirements

Threat Category	AI-Enabled Capability	Primary Security Risk	Required Defense Capability
Adaptive Malware	Changes behavior to evade detection	Endpoint/network compromise	Behavioral and anomaly detection
Automated Phishing	Generates personalized deceptive content	Credential theft	AI-assisted content and identity analysis
Adversarial Attacks	Manipulates AI inputs/models	Incorrect security decisions	Adversarially robust models
Autonomous Reconnaissance	Continuously identifies vulnerabilities	Rapid attack preparation	Real-time threat intelligence
AI-Driven Intrusion	Adapts attack paths dynamically	Network compromise	Continuous adaptive monitoring
Data/Model Poisoning	Corrupts training or operational data	Compromised AI decisions	Data validation and model monitoring
Agent Manipulation	Exploits autonomous agent reasoning/actions	Unauthorized autonomous behavior	Agent isolation, verification, and access control

Intelligent Detection Frameworks

The current paradigms trade off a supervised binary classification paradigm where known patterns are well identified with high fidelity- labelled data- and an unsupervised paradigm,

Volume-1, Issue-3, November-2025 ISSN: 3107-4308 Paper ID: ETDT-V1-I3-04

where unknown patterns are crucial and, most importantly, zero-day, as this paradigm is essential for setting behavioural norms (Chhillar et al., 2026). Hybrid models can combine the accuracy of supervised classifiers with the generative capabilities of deep learning models in an unsupervised manner to minimise false responses in high-velocity environments (Mohale & Obagbuwa, 2025; Okoli et al., 2024). In addition, these behavioral analytics models can be used to profile interactions between latent entities and provide higher detection effectiveness for deviations from such patterns that indicate a multi-stage attack pattern (Kurra, 2025).

Adaptive Defense Mechanisms

Automated prevention options now also utilise cloud-native orchestration and dynamic risk scores to enable real-time enforcement across distributed network endpoints (Antwi et al., 2025). These solutions combine federated learning with reinforcement-based agents to adaptively learn and bound micro-segments, cutting off the isolated segment before any lateral movement is possible (Garg et al., 2026; Shah et al., 2025). These adaptive controls also integrate real-time threat intelligence so that the perimeters' defences are continually updated and security policies evolve as indicators of compromise (IoC) are discovered (Kumar, 2026; Prakash et al., 2024). This combination of predictive analysis and automated response enables an architecture that's not just resistant but also proactive in mitigating threats such as zero-day exploits and polymorphic attacks, which have an edge over signature detection techniques.This integration of predictive analysis and automated response enables a more resilient architecture, one that can neutralise zero-day vulnerabilities — and polymorphic threats, which evade signature-based detection — before they matter (Rani et al., 2025).

Autonomous Response Systems

They are also called automation frameworks or orchestration systems, which streamline and automate the triage, correlation, and prioritisation of security alerts, enabling normative decision-making modules to self-propagate containment or remediation processes (Erigha et al., 2025). By streamlining centralized orchestration, these systems bring simplicity and consistency to operational processes while working with a diverse set of tools, all integrated into one platform with established security ties. Additionally, high-consequence containment actions require an element of "human-in-the-loop" for validation of the AI remediation to ensure that automated logic remains within the organisational risk appetite and is compliant with regulations (Cao et al., 2024; Islam et al., 2019). When human intervention is required, this provides a valuable safety net that reduces the likelihood of faulty automated responses negatively impacting critical business availability (Alevizos & Dekker, 2024).

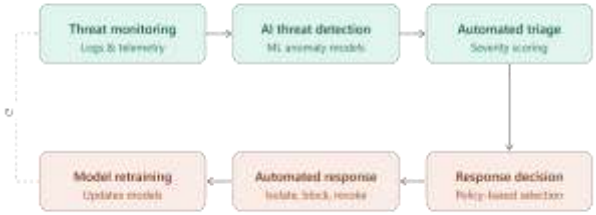


Figure 2. Closed-Loop Autonomous Cyber Threat Detection and Response Cycle

DOI: https://doi.org/10.65713/etdtv1i304

Volume-1, Issue-3, November-2025 ISSN: **3107-4308** *Paper ID:* ***ETDT-V1-I3-04***

Research Gaps and Directions

In addition to advancements, there remains a significant challenge with deep learning models: their decision-making process is opaque and non-interpretable, making it difficult to comprehend autonomous reasoning and to confirm the reliability of autonomous decisions (Reddy & Reddy, 2024). Tthere are currently no techniques to identify any adversarial manipulation of the explanation components generated from the models, which makes it all the more vital to have an inherently explainable model, referred to as a "glass-box" model, or even output consensus from an ensemble of models (Malikussaid & Sutiyo, 2025). In addition to the technical aspects of opacity, other important challenges in the integration of autonomous defence include governance issues and ethical limitations on the responsibility of autonomous agents that must make potentially high-consequence containment decisions (Haryanto et al., 2024).

Evaluation Challenges

There is a lack of common benchmarks in the field to meaningfully compare autonomous solutions and, of course, no specific metric that accurately captures operational efficacy in the real world that is specific to Neuro-Symbolic. These inconsistencies highlight the need for future evaluation criteria that go beyond static performance metrics to assess how well a system works in general, is audited, and can be directly directed (Vyas et al., 2025). To define these metrics, a variety of high fidelity adversarial datasets with diverse configurations is needed to address the complexity of current architectures and distributed systems while preserving privacy-preserving protocols (Pinyi, 2026).

Future Research

The emphasis should be on defining “security-by-design” principles based on adversarial robustness and formal verification, and on the resilience of systems against complex manipulation (Lippi et al., 2025). At the same time, the enhancements of adversarial training, input sanitization, and ensemble modelling remain crucial to fortify the AI models against the evasion techniques (Sarker et al., 2023). In addition, human-in-the-loop strategies have been proposed, in which agents are designed to seek guidance from human experts when facing ethical dilemmas or complex decision points, thereby enhancing the symbiotic relationship and ensuring flexibility in decision-making (Xu et al., 2025). Furthermore, developing knowledge-sharing efforts between academia, industry, and government that will enhance explainable AI solutions to be effective against changing adversarial maneuvering (Rjoub et al., 2023) will be critical. Lastly, investments in privacy-preserving methods like differential privacy and homomorphic encryption, which have the potential to safeguard sensitive training data without jeopardizing non-compliance with international privacy regulations and frameworks, are essential for autonomous defense systems (Familoni, 2024). Further, the research needs to focus on developing an adaptive or "living" regulatory regime that can keep pace with emerging technology and continuously adjust to stay ahead of these autonomous defences, with accountability mechanisms in place (Kulothungan, 2025). In addition to these regulation, the research should be able to enforce systemic "auditability" of actions that allow autonomous agents to record the justification for each action they take to ensure that the answers continue to be lawful and proportional, even as operational tempos exceed humans' cognitive abilities (Lohn et al., 2023). Further, "humanized" AI tools, which can translate the

DOI: https://doi.org/10.65713/etdtv1i304

precise meaning of automated actions in a human-readable manner while respecting context, will ensure that the swift nature of AI aligns with human expert understanding (Achuthan et al., 2024).

References:

1. Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. *Frontiers in Big Data*, *7*, 1497535–1497535. https://doi.org/10.3389/fdata.2024.1497535
2. ADABARA, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Venkateswarlu, M. (2025a). A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense. *F1000Research*, *14*, 843–843. https://doi.org/10.12688/f1000research.169337.1
3. ADABARA, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Venkateswarlu, M. (2025b). Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment. *F1000Research*, *14*, 905–905. https://doi.org/10.12688/f1000research.169927.1
4. Al-E'mari, S., Sanjalawe, Y., & Fataftah, F. (2025). AI-Driven Security Systems and Intelligence Threat Response Using Autonomous Cyber Defense. In *Advances in computational intelligence and robotics book series* (pp. 35–78). IGI Global. https://doi.org/10.4018/979-8-3373-0954-5.ch002
5. Al-E'mari, S., Sanjalawe, Y., Fataftah, F., & Hajjaj, R. Y. (2025). Foundations of Autonomous Cyber Defense Systems. In *Advances in computational intelligence and robotics book series* (pp. 1–34). IGI Global. https://doi.org/10.4018/979-8-3373-0954-5.ch001
6. Alevizos, L., & Dekker, M. (2024). Towards an AI-Enhanced Cyber Threat Intelligence Processing Pipeline. In *Electronics* (Vol. 13, Issue 11, pp. 2021–2021). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/electronics13112021
7. Alhakami, W. (2024). Evaluating modern intrusion detection methods in the face of Gen V multi-vector attacks with fuzzy AHP-TOPSIS. *PLoS ONE*, *19*(5). https://doi.org/10.1371/journal.pone.0302559
8. Antwi, I. K., Akwei, E., Ogundojutimi, O., & Donkor, N. (2025). *AI-Driven Infrastructure Protection Framework for Resilient Enterprise Networks*. 4566–4578. https://doi.org/10.38124/ijisrt/25may2294
9. Bhargava, N. R., & Bommela, N. R. (2025). Mitigating AI-Driven Cyber Threats: A Defense Framework for Securing Autonomous Systems from Intelligent Adversaries. *International Journal of Computer Engineering in Research Trends*, *12*(8), 8–15. https://doi.org/10.22362/ijcert/2025/v12/i8/v12i802
10. Cao, Y., Pokhrel, S. R., Zhu, Y., Doss, R., & Li, G. (2024). Automation and Orchestration of Zero Trust Architecture: Potential Solutions and Challenges. *Machine Intelligence Research*, *21*(2), 294–317. https://doi.org/10.1007/s11633-023-1456-2

11. Chhillar, Dr. K., Tomar, Dr. D., & Shrivastava, Prof. S. (2026). AI Enhanced Intrusion Detection and Prevention Systems (IDS/IPS). *Iconic Research and Engineering Journals*, *9*(9). https://doi.org/10.64388/irev9i9-1714847
12. Dehghantanha, A., Yazdinejad, A., & Parizi, R. M. (2023). *Autonomous Cybersecurity: Evolving Challenges, Emerging Opportunities, and Future Research Trajectories*. 1–10. https://doi.org/10.1145/3689933.3690832
13. Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. (2025). AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Computing Surveys*, *57*(7), 1–36. https://doi.org/10.1145/3716628
14. Devagiri, B. R. (2026a). Agentic AI for Autonomous Cyber Defense. *Journal of Information Systems Engineering & Management*, *11*, 1579–1587. https://doi.org/10.52783/jisem.v11i1s.14327
15. Devagiri, B. R. (2026b). Agentic AI for Autonomous Cyber Defense. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.18635445
16. Devagiri, B. R. (2026c). Agentic AI for Autonomous Cyber Defense. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.18635444
17. Erigha, E. D., Obuse, E., Ayanbode, N., Cadet, E., & Etim, E. D. (2025). Self-Learning autonomous cyber defense agents in AI-empowered security operations. *Computer Science & IT Research Journal*, *6*(8), 475–505. https://doi.org/10.51594/csitrj.v6i8.2011
18. Familoni, B. T. (2024). CYBERSECURITY CHALLENGES IN THE AGE OF AI: THEORETICAL APPROACHES AND PRACTICAL SOLUTIONS. *Computer Science & IT Research Journal*, *5*(3), 703–724. https://doi.org/10.51594/csitrj.v5i3.930
19. Garg, M., Kaza, P. R., Kilaru, A., Basha, S. S., Gunupudi, C., & Kondapalli, K. K. (2026). *Adaptive AI-Driven Threat Intelligence Framework for Proactive Cyber Defense in Dynamic Network Environments*. 1995–2000. https://doi.org/10.1109/icauc68182.2026.11441080
20. Hakim, S. B., Adil, M., Velasquez, A., Xu, S., & Song, H. (2025). Neuro-Symbolic AI for Cybersecurity: State of the Art, Challenges, and Opportunities. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2509.06921