



International Journal of Applied Smart Interdisciplinary Technologies (IJASIT)

Home Page: <https://ijasit.org>

ISSN: 3139-759X

<https://doi.org/10.68104/ijasit.v1.i4.44>

Cyber Defence, Agentic AI, and the Sociotechnical Paradigm: A Survey of AI-Driven Cybersecurity

Rupsingh Puransingh Mathwale

Department of Information Technology, SGGS Institute of Engineering and Technology Nanded, Maharashtra India 431606.

*** Corresponding author.**

E-mail: rpmathwale@sggs.ac.in

ABSTRACT

Article History:

Received Date: 29 August 2026

Revised Date: 06 October 2026

Accepted Date: 11 October 2026

Published Date: 22 October 2026

Keywords:

Autonomous Cyber Defense; Agentic Artificial Intelligence; Sociotechnical Security; Cyber-Physical Systems

Artificial Intelligence (AI) and Large Language Models (LLMs) are reshaping cybersecurity, moving it from signature-based reactive security to agentic, self-evolving, and proactive operations. However, the fast deployment of autonomous agents entails significant security issues, physical-layer attacks in cyber-physical systems, alignment problems, and advanced social engineering risks jeopardizing the stability of the modern sociotechnical infrastructures. This survey paper, which summarizes and analyses the literature, examines the duality of AI in cybersecurity, exploring its implications for strengthening digital defences and increasing the threat landscape for adversaries. We perform a thorough review of 36 cutting-edge research articles, which are organized into six important areas: self-evolving defensive paradigms, LLM-based intelligent agents, reinforcement learning for autonomous decision-making, security of cyber-physical systems, human-machine co-teaming, and agent alignment. Our synthesis shows that multi-agent collaborations greatly shorten incident response times and automate discovery of vulnerabilities but are also highly vulnerable to prompt injection, training data poisoning, and physical-layer spoofing. Moreover, we identify a shift from model patches to model alignment and constitutional safety at runtime as a prerequisite for securing these systems. We believe that resilient digital defence must also be considered a sociotechnical problem, as there must be agents, autonomous systems, that are governed by other agents, aligned and constitutionally bound, and are able to collaborate, rather than operate completely separately, with human operators.

1. Introduction

In the modern world of cybersecurity, attacks have become more common and more complex in numerous ways, such as massive data breaches, coordinated ransomware attacks, worm infections, and the quick exploitation of newly published software vulnerabilities. Traditional enterprise network defenses were based on a deterministic approach, which consisted of perimeter security, static firewall policies, and signature-based Intrusion Detection Systems (IDS). Traditional approaches relied on manually created and inflexible rulesets that were tailored to known and predictable threat patterns. These non-adaptive controls are effective against early, moderately automated attacks but have severe shortcomings against modern, polymorphic, and multi-stage targeted attacks. Traditional security systems are not effective because cyber-attacks are not limited to one

system, vector, or layer of security. Today's threats use machine learning and automation to adapt their exploits on the fly, move across networks at machine speed, and perform sophisticated, context-dependent movements or "living-off-the-land" actions (spear-phishing, etc.). These attacks are constantly changing due to the environmental feedback and defensive countermeasures, so the heuristic is ineffective in identifying the latent indicators spread across various administrative domains.

In addition, a human analyst in traditional Security Operations Centers (SOCs) is always overwhelmed with hundreds of disconnected alerts, which need to be manually validated, prioritized, and mitigated. This "manual triage" has a very high latency and can give the agile attackers enough time for building a long-term persistence or a lateral move before the actual actions of containment.

These challenges have led to a qualitative shift to self-evolving cyber defense. This paradigm revolutionizes digital defense from patch-based, periodic solutions to closed-loop solutions. Near real-time, self-evolving systems that are tightly coupled between sensing, analysis, decision-making, and execution are governed by a continuous autonomic feedback mechanism. These systems are not solely defensive; instead, they continually adapt their strategies based on non-threatening and threatening environmental actors. A self-evolving cyber defense system, based on autonomic computing and control theory, has four key characteristics that ensure temporal robustness and systemic resilience:

Self-Adaptation: Adaptively tuning the detection thresholds, feature representations, and mitigations to overcome concept drift and changing adversarial maneuvers.

Self-Learning: Incorporating continuous online learning, reinforcement learning, or meta-learning directly within the production control loop, thus absorbing new attack patterns without relying on pre-recorded, labeled offline data.

Self-Healing: Auto-containment of incidents (e.g., isolate affected hosts or revoke access tokens) and auto-reconfiguration of services on the fly to reduce Mean Time to Respond (MTTR) and prevent lateral spread.

Self-Optimization: Optimizing security in an effective manner while considering resource limitations, including compute, latency, and network bandwidth, in a principled and mathematically sound way.

1.1 LLMs as Threat Enablers vs. Defensive Tools

To capture the dual-use nature of Transformer-based models, standardized scoping reviews such as those guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) scoping review extension can be carried out.

Adversarial Misuse: Offensive: LLMs as force multipliers reduce the technical hurdles in conducting very advanced attacks. They generate realistic spear-phishing lures, assemble polymorphic malware, and write automated zero-day exploit payloads, among other capabilities, and they are easily manipulated by the adversary.

Defensive Acceleration: LLMs power up defensive operations to a great degree. They can read and understand unstructured text and code blocks and automatically extract cyber threat intelligence (CTI), summarize lengthy reports, provide human analysts with suggestions, detect vulnerabilities at the code level, and create precise remediation patches.

1.2 The Autonomy Gradient: From Passive Models to Agentic AI

Passive LLM deployments are clearly differentiated from autonomous, goal-oriented agentic AI in the literature. Passive LLMs are reactive and output natural language responses to static inputs in the form of prompts. Agentic AI, on the other hand, is a shift to systems that feature goal-driven reasoning, long-term memory, and the power to manipulate digital or physical spaces, using tools, APIs, and teamwork between multiple agents.

Agentic systems are general-purpose automation systems that come with tooling ecosystems that enable chaining of reasoning steps, long-term memory, and external API integration. The Observe-Think-Act (OTA) cycle is at the root of all these systems. The agent keeps observing its surroundings, reasoning upon the retrieved context using powerful planning tools (Chain-of-Thought, self-critique, etc.), and it takes actions while it keeps updating its internal state given feedback from the environment.

Modular multi-agent architectures are used for complex security operations. These systems perform parallel processing of different layers of threats, where each layer is processed by a specific domain-oriented agent (e.g., dedicated email, log, and network scanners). A central recommendation system matches up the

independent threat proofs, recognizing stealthy and multi-vector attacks that are not being detected by isolated sensors. In high-coupling scenarios, communication collapse or turn instability is avoided by explicitly decomposing tasks, monitoring them, and terminating them once they have been completed, as in leader-driven frameworks, which boost task completion rates over non-leader-driven groups.

1.3 Cyber-Physical and Embodied Agent Security

When integrating agentic AI and machine learning into the physical layer of the network, there are new physical-layer constraints. In safety-critical applications such as unmanned aerial vehicles (UAVs), connected and autonomous vehicle (CAV) platoons, robotic systems, and heavy manufacturing, manipulation of algorithms or network disruption can lead to severe physical consequences—and fatalities.

These attacks have been studied in the literature using specialized technical approaches, including deep reinforcement learning (DRL) for making dynamic sequential decisions, generative models for addressing non-stationary environments, and federated learning models to ensure data privacy across a distributed network of robotic nodes. They protect physically embodied agents from the physical-layer vulnerabilities, including DoS attacks, GPS spoofing, and kinematics manipulation, and lead to optimal energy saving and low-latency communication.

1.4 The Sociotechnical Paradigm & the "Human Firewall"

One recurring theme of the past several years of scholarship is that the human element cannot be ignored with the use of a purely technological security solution. AI is a key socio-technical infrastructure embedded in organizational processes, posing risks to the cognitive and behavioral aspects of human actors.

Human Firewall shifts the perspective of human operators from being merely passive failures to active agents of defense in the hybrid socio-technical systems. Human thought is the target of advanced social engineering tactics by adversaries, including:

Overreliance on AI decision support systems (automation bias) and the loss of critical thinking and situational awareness.

Psychological Grooming: The gradual process of developing parasocial connections with conversational agents, which gradually de-conditions rational defenses and leads people to fall into relationship-oriented exploitation.

Agent-Based Modeling (ABM) is used to quantify and mitigate those behavioral vulnerabilities. Multi-turn dialogue scams are simulated in synchronous multi-turn dialogue with dynamic psychological states of human agents (trust, vigilance, stress, workload, and digital financial literacy), and victim-behavior trajectories are modeled. When combined with User and Entity Behavior Analytics (UEBA), ABM allows an organization to use "what-if" scenarios with insiders, identify behavioral drift signatures (e.g., unusually quick decision-making and decreased verification rates), and create them.

1.5 The Alignment Frontier: Constitutional Autonomy

When training phase alignment becomes inadequate, like in conventional training phase alignment methods such as Reinforcement Learning from Human Feedback (RLHF) and standard Constitutional AI, it is an indication of the need for long-horizon agents. These static training-time approaches are not aligned when faced with distribution shift, reward hacking, or value drift in complex, open-ended environments.

The literature introduces the concept of constitutional autonomy to address alignment invariants across the system's operational lifecycle. This paradigm of value alignment is expanded to runtime enforcement. Technically, this can be done by making a modification to the transformer attention mechanism, which is called Constitutional Attention. Constitutional Attention is a method that uses a carefully selected and weighted set of ethical principles as an embedding to be added as a bias term to the standard query-key-value calculation. This architectural invariant features a mechanism for reducing attention to problematic combinations of tokens at inference time, allowing for autonomous decision-making to be contained within moral and safety bounds when deployed in the real world. A taxonomy of threat vectors against agentic AI systems is given in Table 1, while Table 2 provides a comparative analysis of multi-agent coordination topologies.

1.6 Comparison with Related Literature Surveys

While there have been several recent survey studies exploring specific aspects of AI in the cybersecurity domain, the body of literature is still quite disjointed. Previous prominent reviews have largely not considered either the physical-layer constraints of robots or the cognitive-layer aspects of humans (e.g., Jangirala et al. [14]

or Chhabra et al. [1] respectively). The novelty of this survey is the combination of a single 4-layer paradigm, that links agent reasoning and human decision bias to the cognitive layer. Further it considers software pipelines, SAST and automated incident response to the digital layer, cyber-physical systems and alignment to the runtime constitutional autonomy to the alignment layer, as shown in Table 3.

Table 1

Taxonomy of Threat Vectors Targetable Against Agentic AI System

Threat Vector	Attack Mechanism	Security Implications	Primary Mitigations	Reference
Direct Prompt Injection	Uses the data-instruction channel for execution of rogue system instructions	Security guardrails bypassed, system breached, necessary instructions carried out.	Boundaries of Dual-interpreter, Input Filtering, System Instruction Isolation.	[1], [2]
Indirect Prompt Injection	Delivers malicious code through an external source that is interpreted by the agent	Unauthorized use of tools, data exfiltration and stealth lateral movement inside a network.	Authorized tools, human in the loop, bounded execution.	[1], [3]
Training Data Poisoning	Takes charge of training corpora or fine-tuning datasets and adds backdoors to the training process.	Targeted misclassification, hidden backdoor triggers, and subverts of the classification logic.	Strong data provenance, outlier detection, federated adversarial training.	[4], [5]
Tool and API Abuse	Uses agent permissions to perform unauthorized transactions on agents through digital tools or APIs.	Modification of databases without authorization, privilege escalation, scanning of external networks.	Set of principles of selective privilege, strict API input cleaning, inspection of runtime tools.	[1], [6]
Denial of Agent Service (DoAS)	Times prompts with infinite loops or computationally expensive operations to minimize LLM API tokens and resources.	Denial of service, high latency overheads, lack of funds to operate.	Execution timeout boundaries, cognitive cycle depth limits, and token rate control.	[1], [7]

Table 2

Comparative Analysis of Multi-Agent Collaboration and Coordination Frameworks

Topology Type	Communication Strategy	Overhead Complexity	Convergence Guarantees	Primary Security / Resiliency Use Case	Reference
---------------	------------------------	---------------------	------------------------	--	-----------

Leaderless Group Chat	Peer-to-peer broadcast.	High message volume	$(O(N^2))$	Low (Prone to loop instability and infinite turns)	Open-ended brainstorming, unstructured code auditing	[8]
Leader-Driven Hierarchical	Star topology: leader schedules, monitors, and terminates tasks.	Low controlled messages)	$(O(N))$	High (Enforces termination markers and consensus)	High coupling, high cohesion incident response, vulnerability discovery	[8], [9]
Decentralized Federated	Local parameter training; global aggregation server.	Medium (Depends on model weights transfer)		Medium (Assumes non-IID data distribution)	Threat detection in Autonomous Vehicles and Distributed Robotics without compromising privacy.	[10], [11]
Pipeline Sequential	/ Directed acyclic graph; sequential handoff.	Low linear count)	$(O(N))$ step	High (Deterministic linear progression)	The structured patch generation and post-processing of Static Application Security Testing (SAST), and vulnerability database updates.	[12], [13]

Table 3
Frameworks Comparative Literature Matrix

Analytical Dimension	Traditional ML/Robotics Surveys	LLM & Agentic Security Surveys	Self-Evolving Defence Reviews	Our Synthesized Survey Paper	Reference
Scope of Autonomy	Special interest on static, offline classifiers, anomaly detection and robotic kinematics	Provides single agent reasoning analysis, application auditing and threat summarizing analysis.	Consider automated cyber operations (ACO) and moving target defence (MTD) without the use of LLM planning.	Full Autonomy Gradient: Generates learning gradient from passive models to goal-directed multi-agent collaboration (e.g., COSMIC)	[1], [4], [5], [14], [15]
Physical-Layer & Cyber-Physical Systems (CPS)	Good emphasis on the networking of nodes and robotic arms	Omitted. The amount of work on physical layer attack is virtually non-existent and hardware limitations are minimal.	These nodes are not included or only treated as the high-level network topology nodes.	Secure MultiUAV Covert Communication and Decentralized Connected Autonomous Vehicle (CAV) Platoon.	[4], [5], [10], [11]
Socio-Technical	When it's a registered variable	Social Engineering Detection	Omitted. Completely uses closed loop	The Human Firewall: Models adaptations of humans as adaptive	[14], [16]

& Human Dynamics	outside the scope of the variable or a static "user permission" variable	(Phishing Classification) at high level.	machine to machine protection	agents, such as cognitive hacking, psychological grooming and real-time voice-phishing simulations.
Adversarial Game Theory & Policies	Provides basic robust training and defence to evasion/poisoning attacks	Basic jailbreak - only prompt-inject jailbreaks and red-teaming are covered	Strategic pathfinding and defence deployment are a focus in simulators.	Active Policy [22], Optimization: Active [23], Moving Target Defence [24] (MTD) using a combination of competitive reinforcement learning (fictitious play) and deep reinforcement learning
Value Alignment & Governance	Omitted. Old fashioned engineering margin constraints are used for safety.	Touches about some aspects of training-phase alignment (RLHF) and some general aspects of policy	Omitted. Performance measures the only criteria used to assess autonomy, not behavioural safety.	Runtime Constitutional [14], Autonomy: Socio-legal [16] structure(s) at runtime (e.g., Constitutional Attention)
Core Synthesis Contribution	The machine learning security is isolated in the middle of the machine.	Isolated to Text based generative security application(s)	Cut off from the mainstream of cyber defence mechanisms	The Unified Paradigm: This combines cognitive, digital, physical and alignment layers in a roadmap.

2. Methodology

2.1 The research design and systematic scoping review framework

This paper follows a Systematic Scoping Review (SSR) framework to document and identify the high-speed convergence of Artificial Intelligence (AI), Large Language Models (LLMs) and autonomous agent technologies with contemporary digital defence and adversarial operations. This is a research design based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) protocols [14].

Given the high number of emerging and rapidly changing papers in this highly complex and heterogeneous field and the interdisciplinary nature of this research, a scoping rather than a narrative review was chosen, as it can be a rigorous, reproducible mapping process. The reviewed literature covers different computing paradigms, such as software engineering (SAST auditing), network operations (incident response and active defence), physical robotics (UAVs and connected AUVs), behavioural science (social engineering and cognitive hacking) and machine learning theory (alignment and constitutional safety).

The main goal of the research is to combine all these disjoint parts of the discipline into a comprehensive, multi-layered architectural taxonomy that captures the upgrading of the defensive capabilities, and the emergence of new security threats enabled by agentic AI.

2.2 The search strategy

The screening of the corpus and inclusion/exclusion criteria are discussed below.

The research corpus was compiled by adopting a multi-stage research strategy on major academic databases such as IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv. First queries were built to use Boolean operators against main themes:

("Agentic AI" OR "Large Language Model" OR "Multi-Agent System")

AND ("Cybersecurity" OR "Intrusion Detection" OR "Vulnerability Discovery" OR "Incident Response")

AND ("Autonomous" OR "Cyber-Physical System" OR "Co-Teaming" OR "Alignment")

From this first extraction, 38 candidate articles are retrieved. A multi-stage screening process was used for theme consistency and academic focus using the following predefined inclusion (I) and exclusion (E) criteria:

I1: The article must contain a technical framework, an empirical evaluation, or an analytical review directly concerning the relationship between AI/machine learning and cybersecurity.

I2: The study should be directed at one of the following aspects of active software defence, network threat mitigation, cyber-physical systems resilience, socio-technical security, or agentic alignment.

E1: Articles that did not include any cybersecurity, vulnerability defence, or adversarial threat mitigation context to analyse general applications of machine learning (e.g., battery energy optimization, physical chemistry modelling) were excluded.

E2: Papers that concentrate on generic software-engineering code assistants but not on specific SAST, security patching, or automated auditing frameworks were not considered.

These filters narrowed the list to two papers about chemical battery management systems, which were removed, leaving a highly curated set of 36 core academic papers for in-depth qualitative and quantitative synthesis. The flow diagram below (Figure 1) shows the visual structure of this selection process (the PRISMA-ScR).

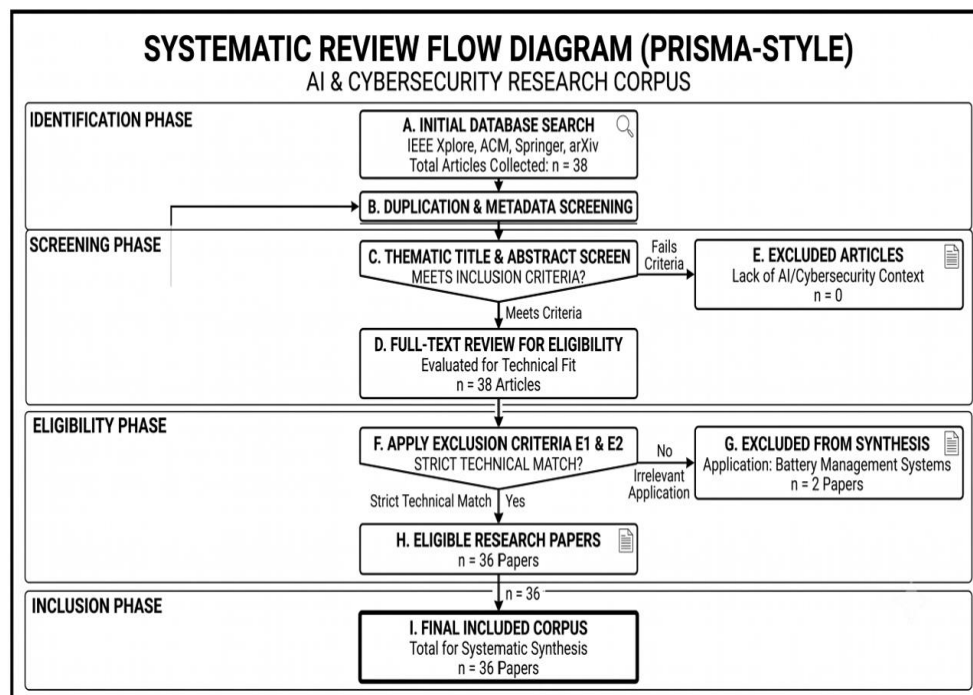


Fig. 1. PRISMA-ScR flow diagram

2.3 Classification Taxonomy and Thematic Synthesis

After finalizing the 36 papers, a grounded thematic analysis methodology was applied to identify the core methodologies, threat vectors, evaluation metrics, and architectural designs from the papers. This coding process resulted in the classification of the papers into six thematic dimensions that correspond to the layers in a contemporary sociotechnical enterprise. We created a systematic processing pipeline to offer a conceptual path through the literature. This pipeline transforms raw scholarly articles into classified architectural features, which will allow the subsequent results to be analysed in a structured, multi-dimensional framework. The taxonomic processing flow is shown in Figure 2. Table 4 illustrates Summary of Core Taxonomic Dimensions, Analysed Variables, and Engineering Methodologies.

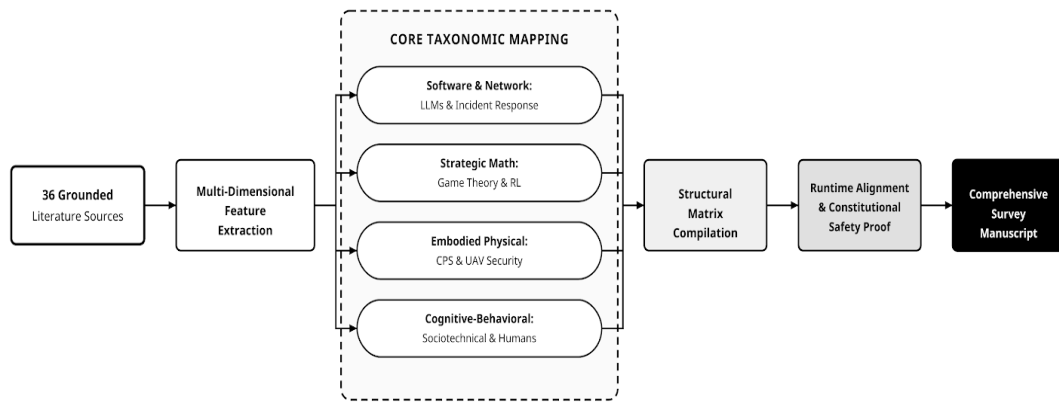


Fig. 2. Taxonomic processing flow

Table 4

Summary of Core Taxonomic Dimensions, Analysed Variables, and Engineering Methodologies

Taxonomic Domain	Key Variables Analysed	Engineering & Analysis Methods	Reference
Self-Evolving Defensive Systems	Detection thresholds, concept drift, Mean Time to Respond (MTTR), continuous anomaly learning rates	MAPE-K autonomous computing loops, online learning classifiers, continuous neural architecture search (NAS).	[15], [18], [19], [20]
LLM-Based Intelligent Agents	False-positive rates, static code patch accuracy, exploit hypothesis efficiency, multi-agent turn-taking dynamics.	Prompt post-processing pipelines, ReAct reasoning, hierarchical leader-driven group chats, vector context retrievers.	[3], [7], [8], [9], [12], [13], [14], [21]
Autonomous Cyber Operations (ACO)	Nash equilibrium exploitability, optimal strategic policies, Moving Target Defence (MTD) performance overheads.	Competitive reinforcement learning, fictitious play within actor-critic, multi-agent DRL, genetic algorithms, supervised mini-batch control.	[22], [23], [24], [25]
Cyber-Physical System (CPS) Security	Trajectory recovery delay, jamming mitigation, passive reflecting beamforming, decentralized training gradient privacy	IRS-assisted secure communications, federated weighted averaging, lightweight language modelling with RL, robust ML control models.	[4], [10], [11], [26], [27], [28]
Sociotechnical & Human Dynamics	Cognitive workload, trust drift, para-social grooming susceptibility, digital financial literacy, fraud detection accuracy.	Agent-Based Modelling (ABM), User and Entity Behaviour Analytics (UEBA), voice synthesis, human-centric co-teaming frameworks.	[6], [17], [29], [30], [31], [32]
Agent Alignment & Governance	Operational-phase alignment validation, query-key-value normative bias calculations, regulatory legal vectors.	Constitutional Attention (runtime attention biasing), policy alignment auditing, decentralized agent registries.	[16], [33], [34], [35], [36]

3. Result & Literature Synthesis

The careful selection of the 36 basic papers and their systematic categorization within specific taxonomies highlights important, multi-dimensional findings from the digital, physical, strategic, and cognitive aspects of

cybersecurity. The section brings together the main findings, architectures, and empirical lessons from the literature.

3.1 LLM-based intelligent agents for software auditing and incident response.

There has been a radical change in the way security is handled from passive language models to autonomous agents that are integrated into tools:

Traditional SAST engines have a high false-positive rate because they lack codebase contexts [13]. This is addressed using a postprocessor based on LLM, such as LSAST and IRIS, which performs deep semantic correlation between repositories to remove non-exploitable bugs, classify vulnerabilities (e.g., memory corruption and injection), and automatically create valid code patches [14].

Offensive frameworks such as VulnBot, RapidPen, and PentestGPT are tools that embody the Observe-Think-Act (OTA) cycle. These systems engage in reasoning loops (e.g., ReAct and Plan-then-Execute) to perform reconnaissance (via network tools); to make a hypothesis about the nature of an exploit; to safely test the network target in isolated virtual environments (e.g., E2B sandboxes); and to create the network payload for the target.

Explainable Security Operations (SOCs): CyberFluxAI shows how to complete the defensive workflow from the streaming telemetry to the guard-railed containment [21]. The system normalizes endpoint, cloud, and network logs; correlates them to the MITRE ATT&CK taxonomy. It generates automated playbooks and mitigation plans that achieved a Mean Time to Respond (MTTR) of 3.30 seconds and a Mean Time to Detect (MTTD) of 1.43 seconds in simulated endpoint and cloud logs in the evaluated streaming telemetry benchmarks of the CyberFluxAI framework [21]. This is just an illustration of the speed at which automated playbooks can be used to contain operations, but in production environments the MTTR/MTTD will also rely on the ingestion latency of telemetry, volume of logs, and the human-in-the-loop validation policies in place. To overcome the “black box” problem, CyberFluxAI incorporates Shapley Additive Explanations (SHAP) and empirical evidence attribution to validate and substantiate each automated answer with specific raw log lines.

Multi-Agent Topologies: Decentralized security structures are based on the deployment of specialized agents that contain domain-specific tools such as dedicated log scanners, email scanners, or registry scanners. Isolated scanners can be easily avoided, but a centralized Contextual Recommendation System can cross-analyse inter-agent risk scores, resulting in an experimental threat detection accuracy of 93.6% and a 41.3% reduction in false positives in correlation benchmarks [12]. These reported metrics shows structured test environments; enterprise-scale setups with real-world setup may exhibit differing baseline false-positives. The COSMIC framework shows that the leader-based group-chat structures, which rely on a Speaker Selector Agent to enforce anti-self-invocation and a Leader Agent to explicitly monitor task completion and provide termination markers, can significantly enhance the convergence of the overall system by 20.2% to 52.6% compared to baseline leaderless chats [8].

3.2 Autonomous Cyber Operations, Game Theory, Reinforcement Learning

Defensive systems must be mathematically based to beat dynamic and adaptive opponents: Defence is a sequential non-stationary game, as is commonly done in Competitive Reinforcement Learning (CRL) environments, including simulated Autonomous Cyber Operations (ACO) environments like the CybORG testbed [22]. Simultaneous training of both Blue (defence) and Red (offence) agents using fictional play in an actor-critic environment is used to approximate minimax optimal policies. These agents learn from previously successful opponents in policies and slowly approach Nash Equilibrium and reduce worst-case exploitability.

Multi-agent Deep Reinforcement Learning (mDRL) for Moving Target Defence (MTD) Deployment: As the topologies are static, the DESOLATER framework uses multi-agent deep reinforcement learning (mDRL) to optimise the resource allocation and deployment schedule of Moving Target Defences [23]. In the field of the Connected Vehicle, deployed in Software-Defined Networks (SDNs), DESOLATER is a Recurrent Neural Network (RNN) approach to learn temporal patterns in host state changes and dynamically assign IP address and port configurations to maximise the attacker's confusion and reduce system packet loss and network overhead.

3.3 Security of Networked Cyber-Physical Systems (CPS) and Autonomous Vehicles

The physical embedding of the autonomous agents in safety-critical nodes, coupled with kinematic constraints, creates tight hardware and kinematic constraints:

Multi-Unmanned Aerial Vehicle (UAV) Networks: Joint optimization of energy usage, transit delay, and signal secrecy are needed in Cooperative Secure UAV Networks [27]. Combining cooperative covert communications (in which friendly UAVs play a jamming noise) with deep reinforcement learning, networks safeguard sensitive telemetry against edge-node eavesdroppers. Moreover, the incorporation of Intelligent Reflecting Surfaces (IRS) in communication paths between UAVs enables agentic beamforming controllers to dynamically control the phase of each element of IRS to reduce the wiretapping signal and enhance the defensive throughput [28].

AV-GPS Spoofing Countermeasures: GPS spoofing attacks against AVs are very dangerous and can be used to alter the target trajectory. To tackle this, the LightLM-RL architecture integrates lightweight transformer embeddings (based on a quantized RoBERTa backbone) with Proximal Policy Optimization (PPO) [26]. Under simulation setup through AV datasets [26], the model has detected semantic GPS anomaly features with 98.57% accuracy and recovery delay of 0.33 seconds.

For fine-grained short-anomaly attacks (e.g., target speed and yaw spoofing), the xFedCAV framework models an intrusion detection system that is decentralized in a cooperative platoon composed of 5 vehicles [10]. Each platoon vehicle builds local anomaly classifiers using LSTM and Transformer models and securely sends the model parameter weights to a central server without sharing raw vehicle telemetries. When a federated weighted aggregation model is employed, it achieves a precision of 98.37% and an F1 score of 98.36% on lead vehicles, which is more than 5% better compared with the stand-alone localized models.

3.4 Socio-Technical Vulnerabilities and the Cognitive Security Layer

When autonomous systems are put into enterprise processes, the human operator is a prime operational target for manipulation of behaviour:

Human Firewall: Instead of simply viewing humans as passive components with inherent weaknesses, the Human Firewall paradigm describes humans as dynamic nodes in a hybrid socio-technical system within high-stress SOC, where operators can fall prey to "automation bias"—the tendency to rely too heavily on AI-generated decision support systems without fully understanding their limitations. Simulations using Computational Agent Based Models (ABM) show that the level of user vigilance reduces over time after successful AI suggestions, resulting in cognitive offloading and high susceptibility to the strategic manipulation of the user by an adversarial agent.

Voice-Phishing Simulations: VishBox [29] includes adaptive multi-turn voice-phishing simulations to help users and to detect behavioural drift. VishBox combines real-time speech synthesis with LLM dialogue agents to mimic socially engineered attacks and monitor the victim in real-time. Such simulated interactions produce granular telemetry data for User and Entity Behaviour Analytics (UEBA) that enables organisations to measure vulnerability factors like stress, trust, workload and more and provide personalised, adaptive security awareness training.

Agentic Payments and Banking Transparency: When financial sectors start using autonomous agents to negotiate, authorise and execute payments, they come with a massive complexity in terms of compliance, legal and operational risks. The EU AI Act and the requirement of "Know-Your-Customer" (KYC) further distribute the legal locus of control between either the users of the platform or the software agents running it and the backends of the banking systems. Integrated banking frameworks employ adversarial robustness testing and feature stability monitors based on a SHAP approach to ensure auditability in such environments: transactional anomaly classifiers keep detecting anomalies with an accuracy of 94.1% even when directly attacked. Table 5 and Figure 3 summarize key performance gains against conventional baselines, to quantitatively demonstrate the effectiveness of the proposed multi-layered framework.

Table 5

Quantitative Comparison Between Conventional Security Baselines and the Proposed Paradigm

Security Dimension	Conventional Baselines	Proposed Framework	Performance Metrics	Reference
Incident Response	Static, rule-based SIEM systems	GenAI-driven orchestration pipelines	Mean Time to Detect (MTTD): 1.43s	[12],[21]

		(CyberFluxAI) with RAG playbook reasoning.	Mean Time to Respond (MTTR): 3.30s	
Multi-Agent Coordination	Unstructured, leaderless peer-to-peer chats; vulnerable to turns non-convergence and infinite loops.	Leader-driven hierarchical coordination	Composite Score Gain: +20.2% to +52.6% over AutoGen/CrewAI Wall-Clock Time Reduction: 50.8%–57.3%	[8]
Autonomous Cyber Operations	Static Moving Target Defense (MTD) schedules; fixed heuristic firewalls.	Competitive Reinforcement Learning	Exploitability Metric: Reaches Nash Equilibrium MTD Overhead: Maximizes attacker entropy while preventing bandwidth degradation	[23]
Autonomous Vehicle GPS Spoofing	Standard Q-Learning / Deep Q-Network (DQN) trajectory tracking models.	Quantized RoBERTa language embeddings	Detection Accuracy: 98.57% (vs. 90.93% baseline) Mitigation Delay: 0.33 seconds (vs. 0.95s baseline)	[26]
Decentralized Platoon Security	Centralized telemetry aggregation; high single-point-of-failure and privacy risks.	Decentralized Federated Agents	Lead Vehicle Precision: 98.37% Lead Vehicle F1-Score: 98.36%	[10]

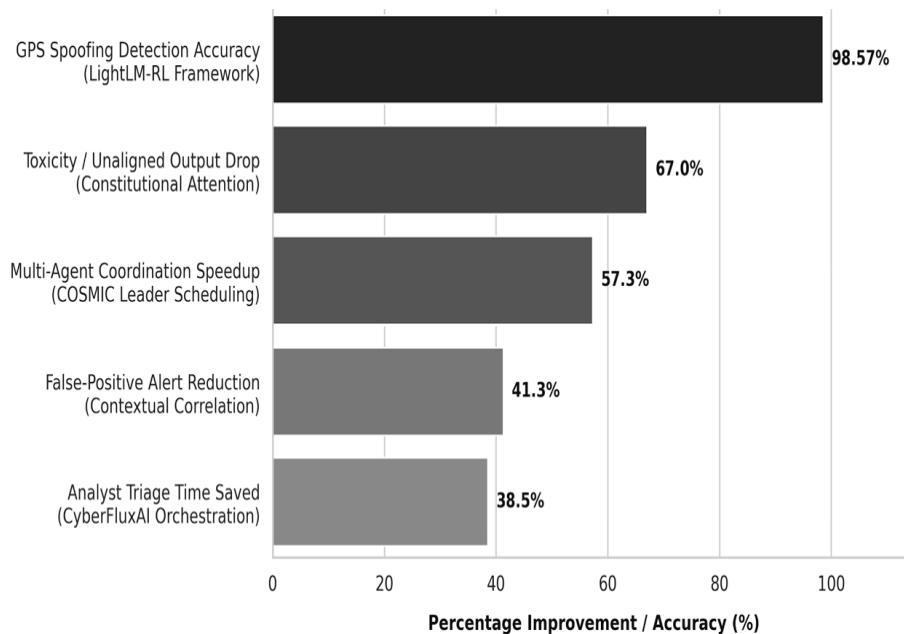


Fig. 3. Quantitative performance improvements

4. Conclusion

Although LLM-powered intelligent agents and multi-agent systems can significantly speed up vulnerability scanning and incident response (like engines CyberFluxAI are able to detect vulnerabilities and respond to them in sub-second time), they also increase the attack surface of the adversary. They pose serious operational weaknesses such as prompt injection, data poisoning, and tool misuse that can be used to break into enterprise networks and physical-layer systems such as multi-UAV networks and connected vehicle platoons (as tested in xFedCAV). Moreover, modern AI operations are sociotechnical infrastructures where security can only be realized by addressing the technological as well as the social aspects in an integrated fashion and not in silos. The human operators are still very susceptible to cognitive hacking, automation bias, and adaptive, multi-turn voice-phishing (as simulated in VishBox). The literature emphasizes this transition towards runtime safety enforcement and constitutional autonomy as a key step to secure this frontier. In architectures such as Constitutional Attention, autonomous systems can keep safety thresholds by hard-coding normative principles into the inference path directly in the model, even with distribution shifts. Finally, to be a resilient digital defense, there must be a sociotechnical perspective on the issue, in which agents aligned with the Constitution, highly auditable, work actively and transparently in partnership with human agents. Future Research Directions and Applications: Autonomous agentic AI brings great benefits to threat mitigation, but it also brings, new operational challenges at the digital, physical and cognitive levels of defence. There are four core technical areas that needs to be addressed to foster resilient and trusted digital ecosystems: Runtime Constitutional Governance: Shift from off-line to static alignments to token-level Constitutional Attention and real-time safety validation (RTSV) loops, to enforce security invariants at runtime. Zero-Trust Inter-Agent Protocols: Implement authenticated Agent2Agent (A2A) communication and least-privilege tool execution sandboxes in Model Context Protocol (MCP) environments. Multimodal & Streaming Telemetry: Support high-throughput pipeline of multimodal inputs (binary payloads, visual attachments) and logs using memory-bounded quantile sketches. Resilient CPS & Adaptive Co-Teaming: Federated Adversarial Training for Edge-constrained Cyber-Physical Systems (UAV, Auton. Vehicle) and Cognitive Guardrails to Counter Human Automation Bias in SOCs.

References

- [1]. Chhabra, S. Datta, S. K. Nahin, and P. Mohapatra, "Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges," *IEEE Access*, vol. 14, pp. 49455–49482, 2026, <https://doi.org/10.1109/ACCESS.2026.3675554>
- [2]. P. Landeta-López, J. M. García, C. Guevara-Vega, and A. Ruiz-Cortés, "LLMs applied to web scraping and web crawling: a systematic review," *Computing*, vol. 108, no. 6, p. 78, Jun. 2026, <https://doi.org/10.1007/s00607-026-01666-5>.
- [3]. R. Özbay, M. Çelebi, and U. Yavanoğlu, "The AI-Cybersecurity Nexus: How Large Language Models are Reshaping Threat Intelligence and Digital Defense," *IEEE Access*, vol. 14, pp. 15558–15587, 2026, <https://doi.org/10.1109/ACCESS.2026.3658308>.
- [4]. S. Neupane et al., "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities," *IEEE Access*, vol. 12, pp. 22072–22097, 2024, <https://doi.org/10.1109/ACCESS.2024.3363657>
- [5]. F. O. Olowononi, D. B. Rawat, and C. Liu, "Resilient Machine Learning for Networked Cyber Physical Systems: A Survey for Machine Learning Security to Securing Machine Learning for CPS," *IEEE Commun. Surv. Tutorials*, vol. 23, no. 1, pp. 524–552, 2021, <https://doi.org/10.1109/COMST.2020.3036778>.
- [6]. D. R. Amariles, D. Charlotin, and L. He-Guelton, "AI Agents in Payments: Applications, Risks and Regulations," *Eur. j. risk regul.*, pp. 1–24, May 2026, <https://doi.org/10.1017/err.2026.10103>
- [7]. Zelenskiy, Leonid, and Andrey Sadovykh. "Automating Cybersecurity Threat Analysis with LLMs." *ACM SIGAda Ada Letters* 45, no. 2 (2025): 88-92, <https://doi.org/10.1145/3821489.3821508>

- [8]. V. Vashishtha, S. Pandey, Y. R. Singh, A. Bahga, and V. K. Madiseti, "COSMIC: Leader-Driven Context-Oriented Collaboration Between Agents," *IEEE Trans. Comput. Soc. Syst.*, vol. 13, no. 3, pp. 3262–3287, Jun. 2026, <https://doi.org/10.1109/TCSS.2026.3661839>
- [9]. A. Alshehri, "Developing a multi-layer agent framework to enhance AI-generated educational questions for cybersecurity," *J. Umm Al-Qura Univ. Eng.Archit.*, vol. 16, no. 4, pp. 1045–1056, Dec. 2025, <https://doi.org/10.1007/s43995-025-00136-x>
- [10]. G. Lin, S. Qian, and Z. H. Khattak, "xFedCAV: Cyberattacks on Leader and Followers in Automated Vehicles With Cooperative Platoons Using Federated Agents," *IEEE Open J. Intell. Transp. Syst.*, vol. 6, pp. 898–914, 2025, <https://doi.org/10.1109/OJITS.2025.3581617>
- [11]. M. Rahmati, "Federated learning for privacy-preserving AI in human–robot collaboration for smart manufacturing," *Journal of Intelligent Manufacturing and Special Equipment*, vol. 6, no. 2, pp. 210–224, Sep. 2025, <https://doi.org/10.1108/JIMSE-03-2025-0003>
- [12]. Y. Hmimou, M. Tabaa, A. Khiat, and Z. Hidila, "A Multi-Agent System for Cybersecurity Threat Detection and Correlation Using Large Language Models," *IEEE Access*, vol. 13, pp. 150199–150215, 2025, <https://doi.org/10.1109/ACCESS.2025.3602681>
- [13]. R. Addis et al., "LLM-Based Intelligent Agents for Cybersecurity: A Tutorial and Survey of Automated Vulnerability Discovery," *IEEE Access*, vol. 14, pp. 100884–100917, 2026, <https://doi.org/10.1109/ACCESS.2026.3707573>
- [14]. S. Jangirala, V. Gupta, and A. Mandal, "Large Language Models in Cybersecurity: A PRISMA-ScR-Guided Scoping Review of Threats, Defenses, and Emerging Applications," *IEEE Access*, vol. 14, pp. 117961–117990, 2026, <https://doi.org/10.1109/ACCESS.2026.3719163>
- [15]. B. A. Alli et al., "Self-evolving cyber defense: an analytical review of AI-driven autonomous and adversarial systems," *J Comput Virol Hack Tech*, vol. 22, no. 1, p. 59, Jul. 2026, <https://doi.org/10.1007/s11416-026-00636-x>
- [16]. W. Torgbi Agbemabiese, "Toward Constitutional Autonomy in AI Systems: A Theoretical Framework for Aligned Agentic Intelligence," *IEEE Access*, vol. 14, pp. 11385–11402, 2026, <https://doi.org/10.1109/ACCESS.2026.3654907>
- [17]. S. Beg, "The human firewall: modeling AI-driven influence and human vulnerability in hybrid ecosystems," *AI & Soc*, vol. 41, no. 7, pp. 7019–7043, Jul. 2026, <https://doi.org/10.1007/s00146-026-02985-z>
- [18]. S. Islam, D. Javeed, M. S. Saeed, P. Kumar, A. Jolfaei, and A. K. M. N. Islam, "Generative AI and Cognitive Computing-Driven Intrusion Detection System in Industrial CPS," *Cogn Comput*, vol. 16, no. 5, pp. 2611–2625, Sep. 2024, <https://doi.org/10.1007/s12559-024-10309-w>
- [19]. M. Shahawy, E. Benkhelifa, and D. White, "Exploring the Intersection Between Neural Architecture Search and Continual Learning," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 36, no. 7, pp. 11776–11792, Jul. 2025, <https://doi.org/10.1109/TNNLS.2024.3453973>
- [20]. A. Foundjem, L. Nganyewou Tidjon, L. Da Silva, and F. Khomh, "Multi-Agent AI Framework for Threat Mitigation and Resilience in Machine Learning Systems," *ACM Trans. Softw. Eng. Methodol.*, p. 3780095, Jan. 2026, <https://doi.org/10.1145/3780095>
- [21]. S. O. Albeez, P. S. Aswin Raaj, M. Nabeel, R. Kanniga Devi, and A. Kaja Mohideen, "CyberFluxAI: A GenAI-Powered System for Rapid and Explainable Cybersecurity Incident Response," *IEEE Access*, vol. 14, pp. 71525–71553, 2026, <https://doi.org/10.1109/ACCESS.2026.3688669>
- [22]. G. Mcdonald, L. Li, and R. A. Mallah, "Finding the Optimal Security Policies for Autonomous Cyber Operations With Competitive Reinforcement Learning," *IEEE Access*, vol. 12, pp. 120292–120305, 2024, <https://doi.org/10.1109/ACCESS.2024.3446310>
- [23]. S. Yoon, J.-H. Cho, D. S. Kim, T. J. Moore, F. Free-Nelson, and H. Lim, "DESOLATER: Deep Reinforcement Learning-Based Resource Allocation and Moving Target Defense Deployment Framework," *IEEE Access*, vol. 9, pp. 70700–70714, 2021, <https://doi.org/10.1109/ACCESS.2021.3076599>
- [24]. X. Cai et al., "Stability Analysis of Networked Control Systems Under DoS Attacks and Security Controller Design With Mini-Batch Machine Learning Supervision," *IEEE Trans.Inform.Forensic Secur.*, vol. 19, pp. 3857–3865, 2024, <https://doi.org/10.1109/TIFS.2023.3347889>

- [25]. A. A. J. Al-Hchaimi, M. A. Al-Shareeda, A. T. Azar, and W. El-Shafai, "Reinforcement Guided Genetic Algorithm for Application Mapping in Network-on-Chip Architectures: Toward Transparent and Efficient MPSoC Scheduling," *IEEE Access*, vol. 13, pp. 177520–177536, 2025, <https://doi.org/10.1109/ACCESS.2025.3616179>
- [26]. H. E. Alami and D. B. Rawat, "RL-Enabled Lightweight Language Models for GPS Spoofing Defense in Autonomous Vehicles," *IEEE Trans. Inform. Forensic Secur.*, vol. 21, pp. 1424–1437, 2026, <https://doi.org/10.1109/TIFS.2026.3652406>
- [27]. X. Zhou, H. Guo, B. Mao, Y. Xun, and N. Kato, "Cooperative Covert Communication and Task Offloading for AI-Empowered Multi-UAV Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 12, pp. 9979–9991, 2026, <https://doi.org/10.1109/TCCN.2026.3710515>
- [28]. W. Xie et al., "IRS-Assisted UAV Secure Communication via Agentic AI," *IEEE Trans. Cogn. Commun. Netw.*, vol. 12, pp. 10311–10326, 2026, <https://doi.org/10.1109/TCCN.2026.3714068>
- [29]. Y. Yang et al., "VishBox: An AI-Agent-Based Adaptive Voice Phishing Simulation Framework for Cybersecurity Education," *IEEE Access*, vol. 14, pp. 39672–39686, 2026, <https://doi.org/10.1109/ACCESS.2026.3667823>
- [30]. M. Albanese et al., "Towards AI-Driven Human-Machine Co-Teaming for Adaptive and Agile Cyber Security Operation Centers," *ACM Trans. Internet Technol.*, p. 3819819, Jun. 2026, <https://doi.org/10.1145/3819819>
- [31]. Hildebrandt, Michael, Shuchisnigdha Deb, Xiaoyun Yin, Jackie Cha, and Heejin Jeong. "Augmented Cognition Meets AI: Enhancing Human Performance with Real-Time, Adaptive, and Trustworthy Intelligence." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 69, no. 1, pp. 10–14. Sage CA: Los Angeles, CA: SAGE Publications, 2025, <https://doi.org/10.1177/10711813251367751>
- [32]. D. Keshavamurthy, M. Kumar, G. Tsaramirsis, and F. Oroumchian, "An AI -Based Framework for Secure and Transparent Banking: Integrating Adversarial Robustness, Interpretability, and Organizational Modeling," *Security and Privacy*, vol. 9, no. 1, p. e70153, Jan. 2026, <https://doi.org/10.1002/spy2.70153>
- [33]. S. Dean, T. K. Gilbert, N. Lambert, and T. Zick, "Axes for Sociotechnical Inquiry in AI Research," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 62–70, Jun. 2021, <https://doi.org/10.1109/TTS.2021.3074097>
- [34]. Z. Aminiranjbar, J. Tang, Q. Wang, S. Pant, and M. Viswanathan, "DAWN: Designing Distributed Agents in a Worldwide Network," Oct. 28, 2024, <https://doi.org/10.36227/techrxiv.173014446.64604104/v1>
- [35]. F. Jiang, C. Pan, K. Wang, P. Michiardi, O. A. Dobre, and M. Debbah, "From Large AI Models to Agentic AI: A Tutorial on Future Intelligent Communications," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 3507–3540, 2026, <https://doi.org/10.1109/JSAC.2026.3660010>
- [36]. M. Schmitt and P. Koutroumpis, "Cyber Shadows: Neutralizing Security Threats With AI and Targeted Policy Measures," *IEEE Trans. Artif. Intell.*, vol. 6, no. 7, pp. 1697–1705, Jul. 2025, <https://doi.org/10.1109/TAI.2025.3527398>