

Review Article

From Generative AI to Agentic Cyber Defence: A Comprehensive Review of Multi-Agent Artificial Intelligence for Autonomous Cybersecurity



Ali H. Al-Shakarchi^{1,*}  | Huthaifa L. Mohamed¹  | Omar Khazal Siddeq¹ 
| Musaab Raad Abdullah Al-Hamammi²  | Aesar A. Altaee¹  | S M Abdul Mueid³ 

¹ Department of Computer Networks and Internet, College of Information Technology, Ninevah University, Mosul, Iraq

² Department of Software, College of Information Technology, Ninevah University, Mosul, Iraq

³ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, Malaysia

Received: 28/7/2026 | Revised: 20/8/2026 | Accepted: 20/9/2026 | Online: 22/9/2026

KEYWORDS

Agentic Artificial Intelligence

Autonomous Cybersecurity

Cyber Defense

Large Language Models

Generative Artificial Intelligence

ABSTRACT

Artificial intelligence (AI) has become a key enabler of modern cybersecurity by enhancing threat detection, vulnerability assessment, malware analysis, and incident response. The evolution from machine learning (ML) and deep learning (DL) to Large Language Models (LLMs) and generative AI has significantly improved automated security analysis, natural language interaction, and cyber threat intelligence. However, current generative AI systems remain largely reactive, relying on human guidance and offering limited capabilities for autonomous planning, reasoning, memory management, and collaborative decision-making. Recent advances in Agentic Artificial Intelligence (Agentic AI) introduce autonomous intelligent agents that can perceive their environment, use external tools, coordinate with other agents, and execute complex multi-step tasks with minimal human intervention. These capabilities have created new opportunities for automating cybersecurity functions, including Security Operations Centres (SOCs), threat hunting, malware analysis, phishing detection, vulnerability management, digital forensics, and incident response. This paper reviews the evolution from generative AI to Agentic AI and examines the role of multi-agent architectures in autonomous cyber defence. It provides a thorough overview of current trends, a taxonomy of existing methods by agents' architecture, collaboration, reasoning, and application domains, and the main issues, challenges, research directions, and open questions for designing trustworthy, explainable, and robust autonomous cybersecurity systems.

1. INTRODUCTION

The digital revolution over the last decade has brought revolutionary change to how governments, organisations, and critical infrastructure operate. Technologies such as cloud computing, IoT, edge computing, fifth-generation

*Corresponding Author (Email): ali.al-shakarchi@uoninevah.edu.iq

(5G) communication, blockchain, and artificial intelligence have driven a paradigm shift in how connectivity and automation are achieved. Although these technologies have improved efficiency and productivity, they have also increased the attack surface for cyberattacks [1], [2].

The cybersecurity threat environment has therefore changed in both magnitude and complexity. Contemporary cyberattacks do not rely only on conventional malware and signature-based attacks; instead, they use methods such as advanced persistent threats (APTs), ransomware attacks, supply chain attacks, phishing-as-a-service attacks, insider attacks, cloud-native attacks, and AI-driven cyberattacks [3], [4]. Attacks have become adaptive, polymorphic, and evasive, making them difficult to counter with conventional approaches. Moreover, the increasing volume of security logs, network traffic, vulnerability assessments, and threat intelligence has made it impossible for human analysts to process the data, making intelligent automation in cybersecurity imperative [5].

In the past, cybersecurity solutions relied largely on signature-based methods, which involved comparing observed behaviour to attack signatures and/or manually created rules. While this approach works well for detecting previously seen threats, it cannot detect new threats, such as zero-day attacks, new malware variants, and constantly evolving attacks, because it requires regular updates to the signature database [6]. As a result, integrating machine learning methodologies into cybersecurity processes has become more common.

There is significant evidence that machine learning techniques, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and ensemble methods, have shown great promise in intrusion detection, phishing, malware, spam, and financial fraud detection [7], [8]. However, these techniques rely on finding statistical patterns in historical data and detecting malicious activities that differ from past attack patterns. Despite their success, traditional ML algorithms usually depend on feature engineering, well-balanced labelled datasets, and regular updates to adapt to new threats.

Deep learning has subsequently addressed several of these limitations by enabling automatic feature extraction from high-dimensional cybersecurity data. Architectures including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Transformers, and Graph Neural Networks (GNNs) have substantially improved performance across numerous cybersecurity tasks, including network intrusion detection, malware family classification, anomaly detection, phishing detection, botnet identification, and fraud analytics [9]–[11]. Graph-based learning models enable security researchers to capture structural relationships among network entities, communication flows, and behavioural interactions that conventional tabular learning methods struggle to represent [11].

Despite their impressive predictive capabilities, most deep learning models remain highly specialised and task-dependent. These models mostly act as predictive classifiers rather than self-decision-making systems. Additionally, they are often treated as black boxes because their internal mechanisms are hard to understand. Besides, these models do not incorporate external information sources, make long-term plans, or carry out several security-related activities at once. Despite these improvements in detection accuracy, cybersecurity operations remain largely unchanged and have not become autonomous decision-making environments.

Transformer-based Large Language Models (LLMs) have ushered in a new era of artificial intelligence systems that can accurately interpret and produce natural language. Systems like GPT-4, Llama, Claude, Gemini, and other foundation models have shown impressive performance in tasks such as document summarisation, code generation, vulnerability explanation, malware understanding, incident report creation, and cyber threat intelligence analysis [12], [13]. In cybersecurity, LLMs are increasingly used to interpret security alerts, explain malicious code scripts, and produce detection rules and vulnerability descriptions.

The adoption of generative AI in cybersecurity has greatly reduced the time needed to analyse security information, making AI-driven decision support easier to understand. Rather than sifting through thousands of security logs and lengthy vulnerability reports, analysts can now converse with AI agents that summarise technical information, search knowledge, and generate recommendations [13], [14]. This capability has improved many cybersecurity processes, especially in SOC environments.

However, existing generative AI solutions remain reactive. Most existing LLM-based cybersecurity assistants process user requests one by one and cannot develop independent plans for future actions, organise tasks, maintain long-term memory of their activities, or interact independently with different cybersecurity tools. This is why human experts handle most complex cybersecurity analyses [15].

The limitations of generative AI have motivated researchers to investigate a new generation of intelligent systems known as Agentic Artificial Intelligence (Agentic AI). Unlike traditional conversational models, Agentic AI combines reasoning, planning, memory, environmental awareness, tool utilisation, and autonomous decision-making within coordinated software agents capable of solving complex multi-step problems. Rather than responding to isolated prompts, these agents pursue predefined objectives, adapt dynamically to environmental changes, collaborate with other specialised agents, and continuously refine their strategies based on newly acquired information [16], [17].

Recent advances in multi-agent architecture show that distributing cybersecurity responsibilities among specialised intelligent agents can substantially improve operational effectiveness. Rather than relying on a single monolithic AI model, multiple autonomous agents can independently perform threat intelligence retrieval, network traffic analysis, vulnerability assessment, malware investigation, forensic reasoning, and incident response, then collaboratively integrate their findings into a unified security decision. This collaborative intelligence better mirrors workflows in human Security Operations Centres and brings us one step closer to self-driving cyber-defence [18].

The evolution of generative AI to agentic AI represents a revolution in intelligent cybersecurity. While generative AI can react to user prompts, agentic AI enables autonomous systems that can plan, reason, manage memory, use tools, and make decisions in multiple steps to accomplish goals with little or no human involvement [16], [19].

This approach is especially useful for modern Security Operations Centres (SOCs) because SOC analysts must handle a large volume of alerts and complex cyber threats. A multi-agent system can divide cybersecurity work into subtasks handled by separate agents responsible for threat intelligence, network monitoring, vulnerability assessment, malware analysis, incident response, and reporting [20]–[22].

Recent advancements in retrieval-augmented generation (RAG), knowledge graphs, and external tool integration have enabled Agentic AI to become more advanced by allowing agents to retrieve current cybersecurity knowledge and make evidence-based decisions [16], [23], [24]. Compared with monolithic AI systems, multi-agent systems offer more flexibility, scalability, and adaptability for the complex task of cyber defence.

Despite their numerous advantages, several challenges remain when working with Agentic AI systems. These include hallucinations, prompt injections, secure communication between agents, privacy issues, explainability, governance, and trustworthy autonomous decision-making [14], [16], [19], [25].

Although recent surveys have examined large language models or autonomous agents independently, relatively few studies have systematically reviewed their integration into collaborative cybersecurity systems. Therefore, this paper provides a focused review of recent developments in Agentic AI for autonomous cyber defence by analysing multi-agent architectures, cybersecurity applications, current challenges, and future research directions.

The main contributions of this review are summarised as follows:

- Reviews recent advances in Agentic AI for cybersecurity published between 2022 and 2026.
- Examines the evolution from generative AI to autonomous multi-agent cyber-defence systems.
- Proposes a taxonomy of multi-agent cybersecurity frameworks based on architecture, collaboration, reasoning capability, and application domain.
- Reviews major cybersecurity applications, including intrusion detection, threat intelligence, malware analysis, phishing detection, incident response, and SOC automation.

- Discusses open research challenges and outlines future directions toward trustworthy and autonomous cyber-defence systems.

The remainder of this paper is organised as follows. Section 2 presents the background and evolution from generative AI to Agentic AI—Section 3 reviews multi-agent architecture and cybersecurity applications. Section 4 compares recent studies, introduces the proposed taxonomy, and discusses current challenges and future research directions. Section 5 concludes the paper.

2. Evolution from Generative AI to Agentic AI

AI has become an important enabler for cybersecurity by enhancing detection, analysis, and response to ever-evolving cyber threats. Supervised ML algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR) were used in the first generations of AI-driven security solutions for classification of attacks based on manually crafted features in network traffic, executable code, and system logs [7]. These methods were much more efficient at identifying novel threats and minimising false positives than classical signature-based approaches to security management [26]. However, their performance depended heavily on feature engineering and regular retraining on labelled data.

Because cybersecurity data accumulates rapidly and cyberattacks evolve quickly, DL methods were quickly adopted. CNN, RNN, LSTM, Transformer, and GNN architectures can automatically extract features and perform better at intrusion detection, malware classification, phishing detection, anomaly detection, and fraud analytics [9], [27]. Graph-based models can identify structural relationships between users, devices, flows, and network entities. For these reasons, graph-based models are well suited to modelling attack behaviour [11]. However, despite their high accuracy, traditional ML and DL models are task-specific and cannot reason, understand context, plan long-term, and autonomously interact with external cybersecurity resources.

The advent of foundation models and Large Language Models (LLMs) was another big achievement in AI-based cybersecurity. Transformer models such as GPT, Llama, Gemini, Claude, and others have shown strong performance in natural language understanding, code generation, vulnerability explanation, malware analysis, cyber threat intelligence summarisation, and automated reporting [12]. Unlike traditional AI models designed for a single task, foundation models are highly knowledgeable because of large-scale pretraining and can be adapted to different cybersecurity applications through prompting or light fine-tuning.

Generative AI can also produce textual explanations, detection rules, security documentation, vulnerability assessments, and software code. Generative AI has been used successfully in malware analysis, phishing, security awareness training, vulnerability management, and SOC support within cybersecurity [13]. Furthermore, Retrieval-Augmented Generation (RAG) helps ensure that foundation models can retrieve updated information from external cybersecurity knowledge sources such as MITRE ATT&CK, the National Vulnerability Database (NVD), and Common Vulnerabilities and Exposures (CVE). This retrieval mechanism improves factual consistency and reduces hallucination [21].

However, even with these improvements, existing AI generators remain largely reactive. Most LLM-based assistants process individual user requests without autonomous goal decomposition, memory persistence, strategic planning, and collaborative decision-making. Therefore, human analysts still need to coordinate the whole investigation process, choose relevant tools, check intermediate results, and take further actions in cybersecurity operations [16].

Recent studies have addressed these shortcomings by proposing Agentic Artificial Intelligence (Agentic AI), an extension of foundational models that includes reasoning, planning, memory management, environment sensing, tool use, and autonomous decision-making. Rather than simply reacting to inputs, Agentic AI agents pursue predetermined goals by breaking tasks into sub-tasks, recalling relevant information, using dedicated cybersecurity tools, assessing each task's results, and continually adjusting their execution based on environmental feedback [16], [19].

One unique feature of Agentic AI is its ability to collaborate and integrate several specialised agents in a collaborative environment. Rather than using one large-scale model, separate agents can handle threat intelligence collection, network traffic inspection, vulnerability detection, malware examination, forensic reasoning, event coordination, and security reporting, which can later be combined into a single decision. This collaborative architecture resembles modern Security Operations Centres, where specialists collaborate and use their specialised skills to examine and address cyber incidents [19]. Moreover, multi-agent systems are easy to scale, maintain, and develop individually because new AI models, cybersecurity tools, and knowledge sources can be integrated as separate agents.

Table 1 compares traditional AI, Generative AI, and Agentic AI in cybersecurity. It shows that intelligent systems have evolved from predictive systems to autonomous systems over time.

Table 1. Comparison of AI Paradigms in Cybersecurity (FM = Foundation Model; Ext. = External knowledge/resources)

Aspect	Traditional	Generative	Agentic
Goal	Predict	Generate	Act
Knowledge	Data	FM	FM + Ext.
Planning	–	Low	High
Memory	–	Context	Persistent
Tools	No	Limited	Extensive
Reasoning	Low	Medium	High
Agents	Single	Single	Multi
Human role	High	Medium	Low
Cyber role	Detect	Assist	Defend

While conventional AI mainly focuses on classification and detection using specialised models, Generative AI expands on these functions through natural language processing and content generation. Agentic AI is even more advanced in application because it uses autonomous planning, memory persistence, tool use, and collaborative reasoning across multiple agents. This makes Agentic AI suitable for implementing complex cybersecurity processes.

Figure 1 shows the evolutionary path of AI systems toward autonomous cyber defence. The process starts with traditional security approaches that use signatures, then moves to machine learning and deep learning models, which gradually improve accuracy by learning from data. With the arrival of foundation models and Generative AI, reasoning abilities and natural language understanding become possible, making intelligent assistance in cybersecurity actions feasible. Building on this, agentic AI and multi-agent systems provide planning, memory, tool usage, and decision-making capabilities, making autonomous cyber-defence systems feasible.

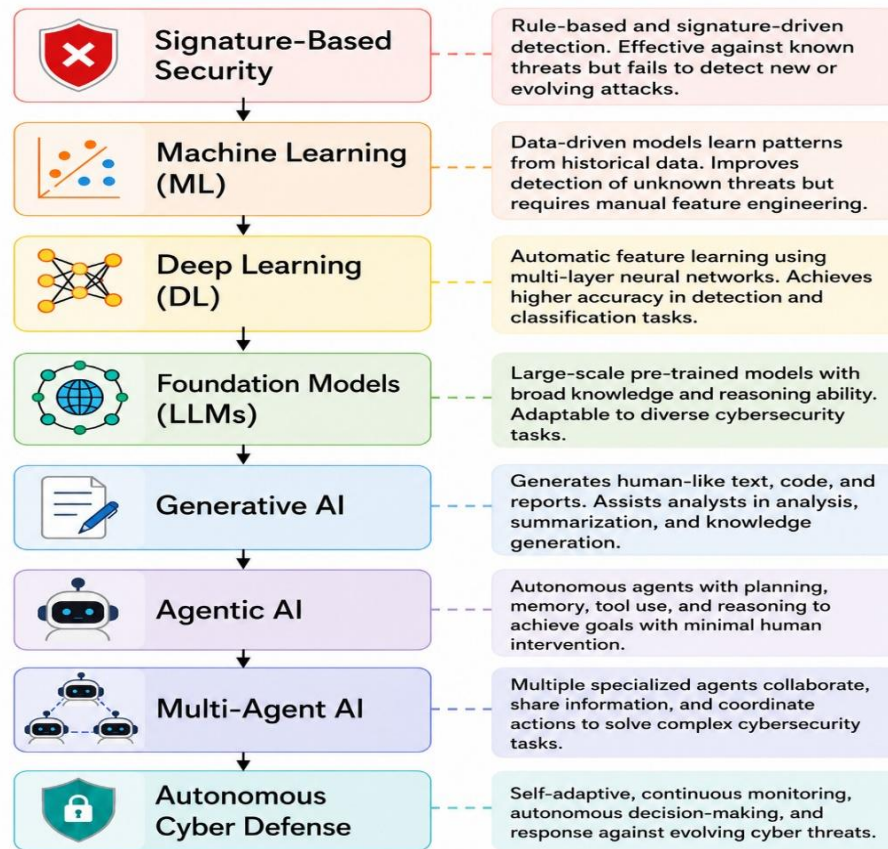


Figure 1. Evolution of AI paradigms in cybersecurity, highlighting the transition from traditional detection methods to autonomous multi-agent cyber-defence systems

Generally, the development of Agentic AI from traditional AI has been characterised by progressively increasing system autonomy. Unlike traditional machine learning algorithms, whose main goal is prediction, deep learning increases representation learning, while generative AI enables language understanding and knowledge generation. Hence, Agentic AI integrates these skills along with autonomous planning, memory, tools and reasoning. As such, AI is shifting from a technology that supports intelligence in decision-making to a participant in performing complex cybersecurity tasks.

2.1 Multi-Agent Architectures for Autonomous Cybersecurity

The shift from generative AI to Agentic AI has transformed the architecture of intelligent cybersecurity systems at its core. Unlike traditional LLM-based assistants that function as standalone conversational models, Agentic AI uses self-sufficient software agents that can reason, plan, manage memory, use tools, and collaborate to make decisions. These agents work together to achieve complex cybersecurity objectives through task decomposition, intermediate communication, and dynamic action adaptation based on environmental feedback [28]–[31]. Collaborative intelligence offers AI systems the opportunity to move beyond decision support toward autonomous cyber defence.

A typical architecture of modern multi-agent cybersecurity systems involves several agents performing different roles in the security process. Rather than delegating all tasks to one AI model, the system divides work among several intelligent components that collaborate to analyse threats, gather knowledge, reason over evidence, and apply necessary defences. This architecture ensures scalability, flexibility, and maintainability, while also enabling the development of separate agents independent of other agents, new AI models, and cybersecurity tools [29], [32].

2.2 Architecture of Agentic Cybersecurity Systems

Despite differences in implementation specifics, most Agentic AI frameworks follow similar architectural pipelines, as shown in Figure 2. The process starts with the Planner agent, which interprets the overall security goal and divides it into subtasks. Examples include malware analysis, intrusion analysis, phishing, vulnerability analysis, and incident response. The Coordinator agent handles the subtasks the Planner agent creates, assigning tasks to agents and coordinating information exchange throughout the process. A Memory agent saves historical observations, intermediate reasoning results, past investigations, and context information for further learning and long-horizon planning. Agents can use their persistent memory to retain their knowledge beyond the context window of typical LLMs [28].

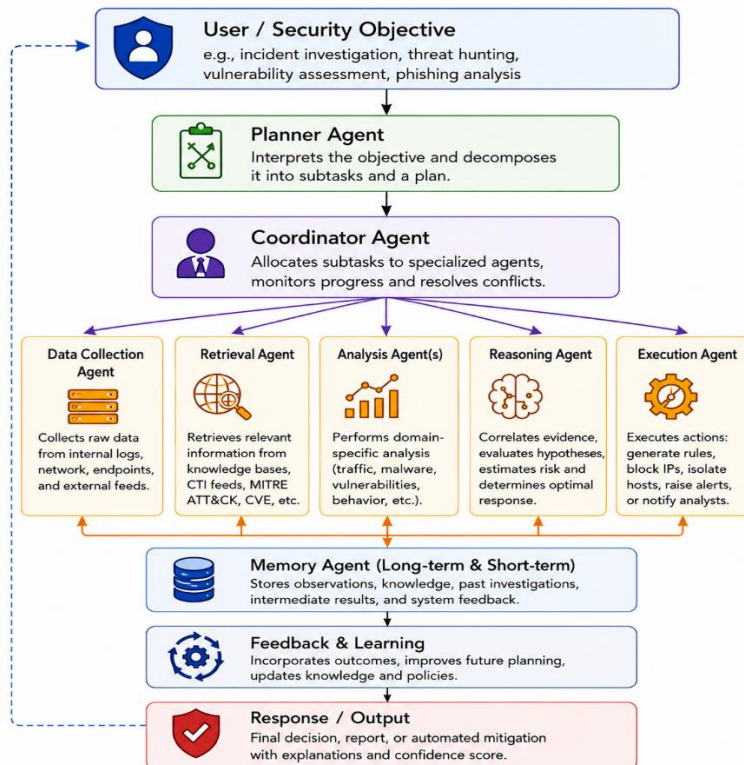


Figure 2. General architecture of a multi-agent Agentic AI framework for autonomous cybersecurity, showing the interaction among planning, coordination, specialised agents, memory, reasoning, execution, and continuous feedback

To improve decision quality, a Retrieval agent continuously gathers necessary information from external cybersecurity knowledge bases, such as MITRE ATT&CK, Common Vulnerabilities and Exposures (CVE), the National Vulnerability Database (NVD), internal knowledge bases, and cyber threat intelligence platforms. Retrieval-Augmented Generation (RAG) combines knowledge from pre-trained sources with updated security information [33].

The system then analyses the extracted information using the Reasoning Agent, where it correlates evidence from various sources, evaluates attack scenarios, determines confidence levels, and identifies mitigation approaches. Lastly, the Execution Agent executes the suggested security actions, including creating detection rules, executing automated responses, and updating firewall policies, among others. The memory module integrates continuous feedback after executing security actions to improve future decision-making [30], [31].

In contrast to conventional systems, this modularity approach decouples perception, reasoning, planning, and execution. This results in greater adaptability since the individual modules can be substituted or improved independently of the rest of the cybersecurity architecture.

2.3 Multi-Agent Collaboration Strategies

The effectiveness of Agentic AI largely depends on how individual agents collaborate to accomplish shared objectives. Existing research generally categorises collaboration strategies into four architectural models.

The single-agent architecture employs one autonomous agent responsible for reasoning, planning, and execution. Although relatively simple to implement, it remains limited in scalability and task specialisation for complex cybersecurity operations.

A hierarchical architecture organises agents into multiple decision levels. High-level planning agents analyse complex goals, and lower agents solve security problems through traffic analysis, malware analysis, or vulnerability assessment. This approach improves coordination and avoids task conflicts, making it well-suited for Security Operation Centres (SOCs) [29].

Collaborative architecture includes communication between agents and constant exchange of reasoning results across the peer agents to make a final security decision. This architecture increases robustness because any individual agent failure has little impact on the investigation process. Lastly, the hybrid architecture combines central planning with distributed task execution. The coordinator agent controls task allocation, while specialised agents solve security problems within a specific domain. The hybrid architecture remains one of the most widely used architectures [31].

2.4 Agentic AI Workflow for Cybersecurity

Although architectures vary, most autonomous cybersecurity solutions follow a common workflow. In the initial phase, security information is gathered from diverse sources such as network traffic, system logs, endpoint telemetry data, vulnerability databases, and external threat intelligence solutions. After processing, the planner agent defines the investigation objective and delegates corresponding tasks to specific agents.

Agents independently execute their assigned analyses while engaging with memory databases and external knowledge bases. Collaborative agents share intermediate results, enabling correlation and reasoning. Using the combined results, the reasoning agent identifies the most likely attack scenario and suggests a response plan. Ultimately, the execution agent executes the recommended defence measures or makes the decision that needs human approval.

This process loop allows autonomous cybersecurity systems to improve their knowledge of emerging cyber threats iteratively and helps human analysts with evidence-based recommendations, as shown in Figure 3.

The workflow in Figure 3 shows the basic difference between traditional AI-assisted cybersecurity solutions and the proposed approach based on autonomous collaboration. Instead of making a single prediction, Agentic AI breaks a security task into a sequence of coordinated subtasks that different agents perform. This architecture lets the task be split among several entities, each focusing on a specific aspect of the problem, such as planning, data search, reasoning, and response execution. The ability to use knowledge from external sources, including threat intelligence databases, vulnerability databases, and cybersecurity frameworks, provides more context and enables evidence-based decisions, unlike standalone language models. Moreover, implementing persistent memory and feedback mechanisms allows the agent to accumulate operational experience and make better decisions in the future. In sum, these features support a shift from reactive automation to proactive, adaptive defence through intelligent collaboration and learning.

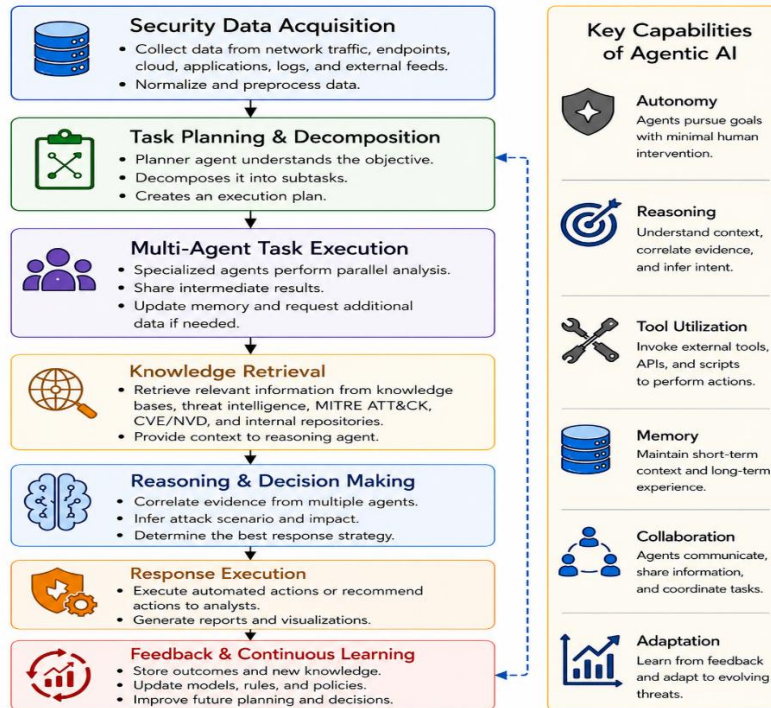


Figure 3. Operational workflow of Agentic AI in cybersecurity, illustrating the sequence from data acquisition and task decomposition to collaborative reasoning, response execution, and continuous learning.

3. Applications of Agentic AI in Cybersecurity

The autonomous reasoning, planning, and collaboration capabilities of Agentic AI have made it applicable across different fields of cybersecurity. Agentic AI is unlike other traditional AI systems, where each of them performs independent prediction; instead, Agentic AI is an integration of several specialised agents which can coordinate observations, reason based on heterogeneous information, and take defence measures with very little human intervention. In this regard, current Security Operations Centres (SOCs) have started to consider multi-agent architectures to automate workflows, minimise reaction time, and increase cyber resiliency [34], [35].

Figure 4 gives an overview of the major cybersecurity applications of the Agentic AI presented in this review. The application domains in Figure 4 show how Agentic AI goes beyond isolated tasks and provides an integrated intelligence layer that can cover the full cybersecurity operations process. Rather than treating threat intelligence, intrusion detection, malware analysis, vulnerability assessment, incident response, digital forensics, and SOC automation as individual operations, it enables them to form an integrated system where different agents constantly share information and make decisions based on the changing security context. For instance, threat intelligence agents can complement intrusion detection with contextual information. At the same time, malware analysis outcomes can support incident response and attack investigations by helping determine the origin, behaviour, and intent of attacks.

At the same time, vulnerability assessment agents can prioritise measures to address threats thanks to data from operational and threat intelligence agents. Hence, Agentic AI serves as an extensible base for creating autonomous systems that can detect and respond to increasingly complex cyber threats, constantly refining their performance through reasoning and learning.

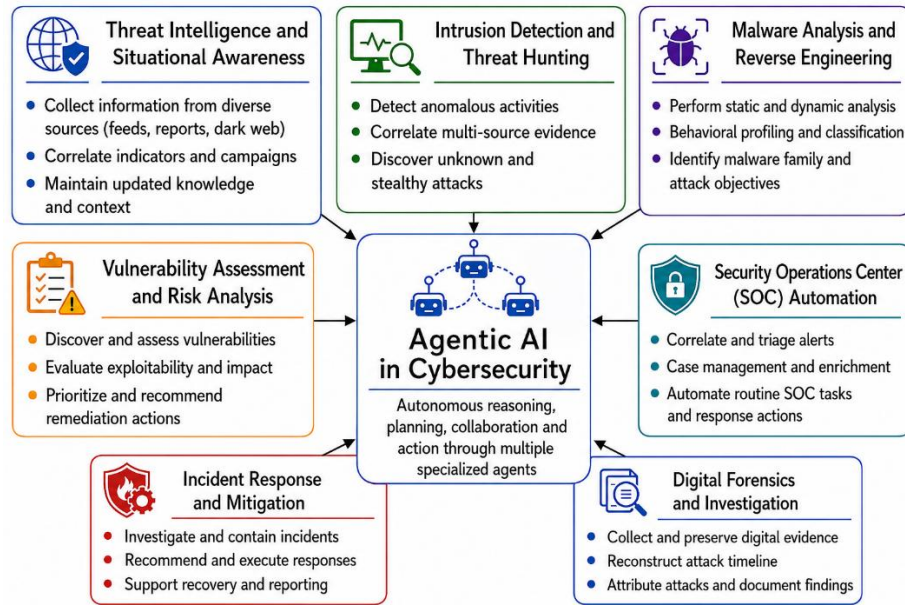


Figure 4. Taxonomy of major applications of Agentic AI in Cybersecurity

3.1 Threat Intelligence and Situational Awareness

The goal of Cyber Threat Intelligence (CTI) is to convert raw security information into useful knowledge that enables proactive defence measures. Agentic AI contributes to CTI by automatically gathering information from heterogeneous sources such as threat intelligence feeds, vulnerability knowledge bases, malware knowledge bases, and security reports. Retrieval agents track information from external knowledge bases, while reasoning agents find correlations between attacks and campaign threats. Retrieval-Augmented Generation (RAG) enhances context by grounding language models in current cybersecurity knowledge [33],[36].

3.2 Intrusion Detection and Threat Hunting

Network Intrusion Detection Systems usually use signature matching or supervised learning models on past network activity. While useful for detecting known attacks, these techniques often struggle to catch more advanced and novel attacks.

Agentic AI helps overcome these problems by having multiple agents analyse network, endpoint, authentication, and behaviour data together. Planner agents coordinate investigation tasks, retrieval agents gather contextual information from external intelligence sources, and reasoning agents correlate evidence from multiple security sensors before generating an overall threat assessment. Such collaborative investigation substantially improves detection capability while reducing analyst workload [34], [37].

3.3 Malware Analysis and Reverse Engineering

Malware analysis requires correlating diverse evidence, including executable behaviour, API calls, memory artefacts, and threat intelligence reports. Agentic AI distributes these tasks among specialised agents responsible for static analysis, dynamic execution monitoring, behavioural profiling, and knowledge retrieval.

Reasoning agents integrate the outputs of individual analysis agents to identify malware families, infer attack objectives, and recommend appropriate mitigation strategies. This collaborative workflow accelerates malware investigation while improving explainability compared with conventional automated malware classifiers [35].

3.4 Vulnerability Assessment and Risk Analysis

Modern enterprise infrastructures continuously generate large volumes of vulnerability information originating from scanners, configuration management systems, software inventories, and external vulnerability databases. Manually prioritising these findings is time-consuming and error-prone.

Autonomous vulnerability analysis is enabled by agentic AI, where the retrieval agent gathers vulnerability data, the reasoning agent analyses exploitability and business impact, and the planning agent schedules remediation actions based on organisational risk. Interaction with external databases such as CVE, NVD, and MITRE ATT&CK continuously updates autonomous agents' situational awareness while supporting risk management [31], [36].

3.5 Security Operations Centre (SOC) Automation

SOCs today suffer from too much alert data that is produced by diverse security monitoring systems. An intelligent orchestration system can be created through Agentic AI to automate various SOC tasks.

In a multi-agent SOC architecture, agents independently perform log analysis, alert correlation, threat intelligence gathering, incident categorisation, evidence summarisation, and recommendations for handling incidents. A coordinator agent combines these intermediate outcomes before generating recommendations or carrying out predefined defensive measures. This architecture greatly reduces analysts' workload and response time [35], [37].

3.6 Incident Response and Digital Forensics

Incident response includes identifying attack sources, understanding adversarial behaviour, limiting the impact on compromised computers, and preserving forensic evidence. Agentic AI supports this task by enabling multiple agents to collaborate throughout the incident response process.

Acquisition agents retrieve forensic artefacts from endpoints and network devices, retrieval agents collect context data from threat intelligence resources, reasoning agents reconstruct attack timelines, and execution agents suggest containment options or initiate automated incident response actions. Memory structures store investigation history and enable later forensic analysis; thus, Agentic AI can be used efficiently in complex, prolonged cyberattacks [35].

Despite the variety of cybersecurity use cases, modern Agentic AI systems share common architectural features. First, they use a modular agent design rather than a single large-scale model, enabling parallel analysis of different types of security data. Secondly, almost all frameworks use Retrieval-Augmented Generation to support factual reasoning based on constantly updated cybersecurity knowledge. Third, collaborative reasoning by multiple agents considerably improves decision-making by correlating evidence collected through separate analysis processes.

However, current approaches have certain limitations. The effectiveness of such methods depends on the correctness of acquired knowledge, effective agent coordination, and adequate protection against hallucinations, prompt injections, and memory poisoning. Therefore, creating autonomous decision-making methods that can be trusted remains an open issue. Table 2 compares several representative studies by their cybersecurity application area, benefits, and limitations.

Table 2. Representative applications of Agentic AI in cybersecurity

Ref.	Focus	Strength	Gap
[19]	General Cyber	Multi-agent reasoning	Low autonomy
[28]	Threat Hunting	Adaptive detection	Coordination
[29]	Playbook Hunting	Automated investigation	Cost
[30]	IDS	Evidence correlation	RAG dependency
[31]	Governance	Ethical design	Immaturity

[32]	Agent Evolution	Automation	Hallucination
[33]	Survey	Prototyping	Standardization
[34]	Multi-Agent Sec.	Threat taxonomy	Oversight
[35]	Autonomous SOC	Governed autonomy	Safety
[36]	Attack-Defense	Threat mapping	Scalability
[37]	Evaluation	Benchmarking	Verification

As shown in Table 2, research on Agentic AI is developing quickly and spans multiple areas of cybersecurity, including threat hunting, intrusion detection, autonomous SOC operations, governance, and security assessment. Although current Agentic AI frameworks improve reasoning, adaptation, and collaboration compared with traditional AI frameworks, several issues remain unresolved. These issues include poor agent coordination, reliance on external knowledge retrieval, scalability problems, hallucination risks, safety issues, and a lack of benchmarking procedures.

4. Comparative Analysis and Proposed Taxonomy

This section is the review's key innovation. Rather than simply presenting an overview of the literature, it synthesises it into a common framework and highlights the latest developments in Agentic AI used in cybersecurity. The suggested taxonomy presents a systematic view of how autonomous intelligent agents are structured, interact and perform various cybersecurity tasks.

4.1 Comparative Analysis of Existing Agentic AI Frameworks

Current agentic AI architectures can be broadly categorised by their architecture design, reasoning skills, collaboration techniques, memory usage, and coupling with external cybersecurity systems. Early designs used a single-agent architecture, where a single large language model interpreted, reasoned, and produced responses for the task provided. While these models showed remarkable natural language understanding, they were not scalable, parallelizable, or robust enough to handle multi-stage cybersecurity tasks [28], [29].

To address these shortcomings, research in the field has developed a multi-agent architecture in which roles are divided among multiple agents. The planner agents break down the security goals into smaller sub-goals; the coordinator agents manage the task assignment and communication process; the retrieval agents gather context information from other knowledge sources; the reasoning agents correlate the evidence; and finally, the execution agents take the defensive action or provide recommendations for human analysts [19], [30], [31].

Another key difference between current frameworks is their use of external cybersecurity knowledge. Traditional LLM-based frameworks depend mainly on pre-training knowledge and may therefore give outdated or erroneous recommendations when faced with new security threats. Modern Agentic AI frameworks use Retrieval-Augmented Generation (RAG) to retrieve relevant knowledge from constantly updated data stores such as MITRE ATT&CK, CVE, NVD, CAPEC, and organisations' cyber threat intelligence platforms [32], [33]. Such frameworks benefit from both retrieved knowledge and LLMs' reasoning, providing higher accuracy in factual claims and lower hallucination rates.

However, collaboration approaches also vary widely in the existing literature. The hierarchical architecture uses centralised planning and distributed execution and hence is especially appropriate for the Security Operations Centre (SOC), where a collaborative decision-making approach is needed. Collaborative architecture uses peer-to-peer communication between independent agents and enables the exchange of partial results from independent analysts before making an overall decision. Hybrid architecture combines centralised orchestration with decentralised reasoning and represents one of the most promising approaches, having been developed relatively recently [19], [30], [34].

Despite these developments, several issues remain unresolved. For instance, existing architectures often rely on task decomposition, inter-agent communication, reliable external knowledge sources, and efficient memory management. Additionally, the lack of standard evaluation criteria, repeatable experimental procedures, and uniform performance metrics makes objective comparisons across Agentic AI frameworks difficult [31], [36].

Overall, the analysed literature shows a clear shift from monolithic AI assistants to more distributed intelligent ecosystems that interact and exchange knowledge continuously in response to changing cybersecurity environments. Table 3 below summarises a comparative study of Agentic AI frameworks for cybersecurity based on their architecture, field of application, main properties, and challenges.

Table 3. Comparative analysis of representative Agentic AI frameworks for cybersecurity

Ref.	Type	Focus	Key Feature	Challenge
[19]	Multi	General Cyber	RAG + Memory	Coordination
[28]	Planner	Threat Hunt	CTI + Planning	Tool dep.
[29]	Multi	Playbook Hunt	LLM + Playbooks	Latency
[30]	Hybrid	IDS	Agent Correlation	Explainability
[31]	Review	Governance	Autonomy + Ethics	Maturity
[32]	Workflow	Agent Evolution	Multi-Agent Pipeline	Cost
[33]	Collaborative	Survey	Use Cases	Trust
[34]	Hybrid	Security	Threat Taxonomy	Scalability
[35]	Ecosystem	Auto. SOC	Tools + Governance	Safety
[36]	Autonomous	Attack-Defense	Threat Mapping	Hallucination
[37]	Agentic	Evaluation	Defense Benchmark	Governance

Table 3 compares some Agentic AI systems by architectural approach, application domain, distinctive properties, and known challenges. The table shows a trend from single-agent architectures to more complex multi-agent approaches with collaboration, memory, and Retrieval-Augmented Generation. These systems greatly enhance automation and adaptability but raise new challenges related to explainability and coordination.

4.2 Proposed Taxonomy of Agentic AI for Cybersecurity

Based on a thorough literature review, this paper proposes a taxonomy for Agentic AI in cybersecurity, as illustrated in Figure 5. Unlike previous classification systems that mainly consider architecture or individual applications, the proposed taxonomy divides existing work into five dimensions.

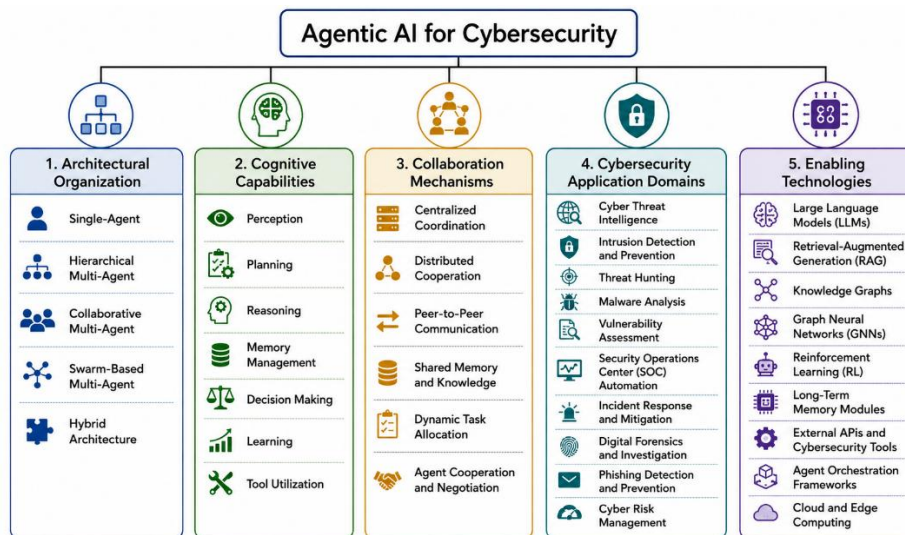


Figure 5. Proposed taxonomy of Agentic AI for cybersecurity

The first is architectural organisation, where systems are classified by how they structure intelligent agents. Studies show progress from single-agent architectures to hierarchical, cooperative, swarm-like, and hybrid multi-agent systems that offer more autonomy, specialisation, and scalability [19], [30].

The second is the capabilities of autonomous agents, including perception, planning, reasoning, memory management, decision-making, learning, and tool use. Agentic AI differs from traditional generative AI in that the latter generates isolated responses, while the former uses autonomous agents with specific capabilities [28], [32].

The third dimension captures collaboration methods, including central coordination, distributed cooperation, peer-to-peer communication, shared memory space, and dynamic task assignment. With collaboration in place, specialised agents can exchange results from their reasoning processes, incorporate heterogeneous security evidence, and solve complex cybersecurity issues that no single model can solve [29], [34].

The fourth dimension denotes the application areas of Agentic AI in cybersecurity. The literature review points to an increased use of collaborative intelligent agents in cyber threat intelligence, intrusion detection, malware analysis, vulnerability assessment, Security Operations Centre automation, incident response, digital forensics, phishing detection, and cyber risk management [30], [33], [36].

Lastly, the fifth dimension includes technologies that enable the development of Agentic AI solutions. These technologies include large language models, retrieval-augmented generation, knowledge graphs, graph neural networks, reinforcement learning, long-term memory modules, cybersecurity toolkits, orchestrators, and APIs. Using these technologies greatly improves reasoning quality, adaptability, transparency, and efficiency [31], [32], [37].

The taxonomy developed above provides a comprehensive conceptual framework that not only summarises previous studies but also lays the groundwork for future development of Agentic AI solutions to cybersecurity challenges.

4.3 Discussion

The comparative analysis of the approaches discussed suggests a fundamental shift from standalone AI-powered cybersecurity tools to collaborative agentic autonomous systems comprising specialised intelligent agents. Rather than fully replacing humans, Agentic AI serves as a smart assistant that can automate routine investigations, correlate heterogeneous evidence, retrieve cybersecurity information, and provide explainable recommendations in complex security incidents [30], [34].

However, multiple issues still prevent the broad use of such technology. Currently available solutions suffer from the problems of hallucination, prompt injection, unreliable external information retrieval, memory poisoning and poor coordination of collaborating agents [31], [35], [36]. Furthermore, evaluation methodology and security benchmarks remain underdeveloped, as existing Agentic AI frameworks lack them.

Hence, future Agentic AI frameworks will need to focus on trustworthy autonomous reasoning, secure multi-agent collaboration, explainable decisions, adaptive learning, and integration with cyber threat intelligence. This will facilitate the development of autonomous cyber-defensive systems that can assist next-generation Security Operation Centers.

5. Conclusion

In summary, to enable the broader use of Agentic AI in cybersecurity, we must address technical, operational, and regulatory challenges that currently limit its practical application of Agentic AI should be solved. While collaborative autonomous agents have improved reasoning, planning, knowledge retrieval, and adaptive decision-making, current models and systems still face problems such as hallucinations, prompt injection, knowledge poisoning, inefficient coordination, lack of explainability, and privacy threats. To build reliable, secure multi-agent architectures, research should focus on improving the reliability of autonomous reasoning, enabling secure communication between agents, developing mechanisms for managing long-term memory, implementing graph-based knowledge representation and Retrieval-Augmented Generation, and establishing standard evaluation methods. Other research directions in this area include improving federated collaboration

among agents, developing human-AI cooperative intelligence, adaptive learning, and post-quantum security. These research areas can help lay the groundwork for reliable, explainable, and fully autonomous cybersecurity systems.

Conflicts of Interest

"The authors declare no conflicts of interest".

Funding

The Authors didn't receive any funding

Acknowledgment

The authors thank Ninevah University for supporting this work.

REFERENCES

- [1] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson, "Threat analysis of IoT networks using artificial neural network intrusion detection system," in Proc. 2016 Int. Symp. Networks, Computers and Communications (ISNCC), Yasmine Hammamet, Tunisia, 2016, pp. 1–6, <https://doi.org/10.1109/ISNCC.2016.7746067>.
- [2] S. Russell and P. Norvig, "A modern, agent-oriented approach to introductory artificial intelligence," ACM SIGART Bull., vol. 6, no. 2, pp. 24–26, Apr. 1995, <https://doi.org/10.1145/201977.201989>.
- [3] S. Mansfield-Devine, "Ransomware: Taking businesses hostage," Network Security, vol. 2016, no. 10, pp. 8–17, Oct. 2016, [https://doi.org/10.1016/S1353-4858\(16\)30096-4](https://doi.org/10.1016/S1353-4858(16)30096-4).
- [4] S. Arora, "Tech turbulence: How CrowdStrike and Microsoft's IT outages disrupted the global economy," SSRN Electron. J., 2024, <https://doi.org/10.2139/ssrn.4902655>.
- [5] S. Mansfield-Devine, "Bad behaviour: Exploiting human weaknesses," Computer Fraud & Security, vol. 2017, no. 1, pp. 17–20, Jan. 2017, [https://doi.org/10.1016/S1361-3723\(17\)30008-8](https://doi.org/10.1016/S1361-3723(17)30008-8).
- [6] W. Stallings, "Physical security essentials," in Computer and Information Security Handbook. Amsterdam, The Netherlands: Elsevier, 2009, pp. 627–643, <https://doi.org/10.1016/B978-0-12-374354-1.00036-4>.
- [7] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," Cybersecurity, vol. 2, no. 1, Jul. 2019, <https://doi.org/10.1186/s42400-019-0038-7>.
- [8] A. M. Albarrak, "An adaptive intrusion detection system for securing the Internet of Medical Things using deep learning," EAI Endorsed Trans. Internet Things, vol. 11, Feb. 2026, <https://doi.org/10.4108/eetiot.10326>.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, <https://doi.org/10.1038/nature14539>.
- [10] D. Akgun, S. Hizal, and U. Cavusoglu, "A new DDoS attacks intrusion detection model based on deep learning for cybersecurity," Computers & Security, vol. 118, Art. no. 102748, Jul. 2022, <https://doi.org/10.1016/j.cose.2022.102748>.
- [11] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4–24, Jan. 2021, <https://doi.org/10.1109/TNNLS.2020.2978386>.
- [12] OpenAI GPT-3.5, "Spor ve teknoloji: Etkileşim ve dönüşümün yolculuğu," in Dijital Çağda Spor Araştırmaları I. Özgür Yayınları, Sep. 2023, <https://doi.org/10.58830/ozgur.pub222.c951>.
- [13] A. Thieme, A. Nori, M. Ghassemi, R. Bommasani, T. O. Andersen, and E. Luger, "Foundation models in healthcare: Opportunities, risks & strategies forward," in Extended Abstracts of the 2023 CHI Conf. Human

- Factors in Computing Systems (CHI EA '23), Hamburg, Germany, 2023, pp. 1–4, <https://doi.org/10.1145/3544549.3583177>.
- [14] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0)—Japanese Translation. Gaithersburg, MD, USA: NIST, 2024, <https://doi.org/10.6028/NIST.AI.100-1.jpn>.
- [15] H. Zhang, P. S. Yu, and J. Zhang, “A systematic survey of text summarisation: From statistical methods to large language models,” *ACM Comput. Surv.*, Apr. 2025, <https://doi.org/10.1145/3731445>.
- [16] L. Wang et al., “A survey on large language model based autonomous agents,” *Front. Comput. Sci.*, vol. 18, no. 6, Mar. 2024, <https://doi.org/10.1007/s11704-024-40231-1>.
- [17] Z. Qin and J. Zhou, “Eva: An LLM-based multilingual voice-agent network for restaurant operations,” in *Proc. 23rd Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, 2025, pp. 3035–3037, <https://doi.org/10.65109/ujp25627>.
- [18] H. Jiang, “Multi-agent large language models for complex text classification,” in *Proc. 2025 3rd Int. Conf. Artificial Intelligence and Automation Control (AIAC)*, 2025, pp. 241–247, <https://doi.org/10.1109/AIAC68175.2025.11332429>.
- [19] P. Jiao, H. Chen, X. Guo, Z. Zhao, D. He, and D. Jin, “A survey on temporal interaction graph representation learning: Progress, challenges, and opportunities,” in *Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2024, pp. 10499–10507, <https://doi.org/10.24963/ijcai.2024/1166>.
- [20] R. Alexander, L. Ma, Z.-L. Dou, Z. Cai, and Y. Huang, “Integrity, confidentiality, and equity: Using inquiry-based labs to help students understand AI and cybersecurity,” *J. Cybersecurity Educ. Res. Pract.*, vol. 2024, no. 1, Oct. 2023, <https://doi.org/10.32727/8.2023.34>.
- [21] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, “Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks,” in *Advances in Neural Information Processing Systems* 36, 2023, pp. 48573–48602, <https://doi.org/10.52202/075280-2109>.
- [22] P. Rafiey and A. Namadchian, “Mapping vulnerability description to MITRE ATT&CK framework by LLM,” *Adv. Artif. Intell. Mach. Learn.*, vol. 5, no. 3, pp. 4379–4396, 2025, <https://doi.org/10.54364/aaiml.2025.53243>.
- [23] National Institute of Standards and Technology, FIPS PUB 58-1997, 1997, <https://doi.org/10.6028/NIST.FIPS.58-1997>.
- [24] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, “A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges,” *Vicinagearth*, vol. 1, no. 1, Oct. 2024, <https://doi.org/10.1007/s44336-024-00009-2>.
- [25] RAND Corporation, “Report associated with DOI 10.7249/CPA2351-1-V3,” RAND Corporation, 2024, <https://doi.org/10.7249/CPA2351-1-V3>.
- [26] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, “Internet of Things security and forensics: Challenges and opportunities,” *Future Gener. Comput. Syst.*, vol. 78, pp. 544–546, Jan. 2018, <https://doi.org/10.1016/j.future.2017.07.060>.
- [27] A. Aldweesh, A. Derhab, and A. Z. Emam, “Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues,” *Knowl.-Based Syst.*, vol. 189, Art. no. 105124, Feb. 2020, <https://doi.org/10.1016/j.knosys.2019.105124>.
- [28] A. Sheth, A. Patel, C. Upadhyay, H. Ragothaman, B. Patil, and S. K. Udayakumar, “Agentic AI for autonomous cyber threat hunting and adaptive defense in dynamic security environments,” in *Proc. 2025 IEEE Int. Conf. Electro Information Technology (eIT)*, 2025, pp. 316–321, <https://doi.org/10.1109/EIT64391.2025.11103697>.
- [29] N. Chen, R. Lin, D. Xie, H. Lin, and S. Chen, “iThelma: An autonomous LLM agent for cyber threat hunting via playbook-driven intelligence,” in *Proc. 2025 IEEE Conf. Communications and Network Security (CNS)*, 2025, pp. 1–6, <https://doi.org/10.1109/CNS66487.2025.11195050>.
- [30] Y. Hmimou, M. Tabaa, A. Khiaat, 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>.
- [31] I. Adabara, B. O. Sadiq, A. N. Shuaibu, Y. I. Danjuma, and M. Venkateswarlu, “A review of agentic AI in cybersecurity: Cognitive autonomy, ethical governance, and quantum-resilient defense,” *F1000Research*, vol. 14, Art. no. 843, Sep. 2025, <https://doi.org/10.12688/f1000research.169337.1>.

- [32] V. Vinay, "The evolution of agentic AI in cybersecurity: From single LLM reasoners to multi-agent systems and autonomous pipelines," in Proc. 2026 IEEE 5th Int. Conf. AI in Cybersecurity (ICAIC), 2026, pp. 1–8, <https://doi.org/10.1109/ICAIC67076.2026.11395809>.
- [33] N. Pachaiyappan, "Agentic commerce: A comprehensive analysis of cybersecurity risks, privacy challenges, and trust mechanisms in autonomous AI-driven marketplaces," in Proc. 2026 IEEE 5th Int. Conf. AI in Cybersecurity (ICAIC), 2026, pp. 1–6, <https://doi.org/10.1109/ICAIC67076.2026.11395759>.
- [34] L. D. Richardson, L. G. Davies, K. Lane, and C. Cronin, "Women's lived experiences of fatness," in Proc. 6th Annu. Int. Weight Stigma Conf., 2018, <https://doi.org/10.31076/2018.p14>.
- [35] H. Alqahtani and P. Ahuja, "Secure autonomous cyber defense with LLM agents: A systematic review of autonomy, tool-augmented reasoning, and governance constraints," Comput. Electr. Eng., vol. 135, Art. no. 111184, Jul. 2026, <https://doi.org/10.1016/j.compeleceng.2026.111184>.
- [36] A. Stavropoulos, "Commentary: Qi, J. et al. Towards trustworthy agentic AI: A comprehensive survey of safety, robustness, privacy, and system security. Acad. AI Appl. 2026, 2, 8260," Academia AI Appl., vol. 2, no. 3, Jul. 2026, <https://doi.org/10.20935/acadaai8435>.
- [37] A. 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>.