



Journal: <https://research-education.com/index.php/ijaie>

Agentic Artificial Intelligence in Cybersecurity: Emerging Risks, Governance Challenges, and Defensive Frameworks

Author

Santosh Kumar Jadala

Cyber Security & Business Analysis Specialist
Independent Researcher

Abstract

Agentic artificial intelligence is emerging as a new class of autonomous AI systems capable of reasoning, planning, tool use, information retrieval, and independent task execution across digital environments. In cybersecurity, these capabilities create major opportunities for threat detection, vulnerability assessment, incident triage, and automated response. However, they also introduce new risks because agentic AI systems may access sensitive data, interact with APIs, execute security-related actions, and operate with limited human supervision. This article examines the cybersecurity risks, governance challenges, and defensive frameworks associated with agentic AI deployment. It highlights key concerns such as prompt injection, unauthorized tool use, data leakage, insecure automation, supply-chain exposure, weak accountability, and insufficient runtime monitoring. Drawing on AI risk management guidance, generative AI governance principles, zero trust architecture, and research on tool-connected language models, the article argues that secure agentic AI adoption requires a layered defensive approach. This includes agent identity control, least privilege access, secure tool governance, prompt and input validation, behavioral monitoring, audit logging, explainability, and human oversight for high-risk actions. The article concludes that agentic AI can strengthen cybersecurity operations, but only when deployed within a transparent, accountable, and continuously monitored governance framework that balances autonomy with organizational control (NIST, 2023, 2024; OWASP Gen AI Security Project, 2025; Schick et al., 2023; Yao et al., 2022).

Keywords: *Agentic AI, cybersecurity, autonomous agents, AI governance, cyber risk, zero trust, prompt injection, AI risk management, defensive frameworks.*

1. Introduction

1.1 Background of the Study

Artificial intelligence has become a central part of modern cybersecurity, moving from rule-based detection systems and traditional machine learning models toward more advanced generative and agentic systems. Traditional AI systems are commonly used to detect patterns, classify malware, identify anomalies, and support automated decision-making within predefined limits. Generative AI has expanded these capabilities by producing text, code, summaries, explanations, and security intelligence outputs. More recently, agentic artificial intelligence has introduced a further shift by enabling AI systems to reason through tasks, retrieve external information, interact with digital tools, call application programming interfaces, and perform multi-step actions with limited human input.

Agentic AI differs from earlier AI systems because it is not limited to passive prediction or content generation. It can pursue a goal, plan a sequence of actions, access external resources, use tools, and revise its behavior based on feedback from the environment. This development is supported by research on retrieval-augmented generation, reasoning-and-acting models, tool-using language models, and API-connected large language models. Retrieval-augmented generation allows language models to access external knowledge sources rather than relying only on stored model parameters (Lewis et al., 2020). ReAct demonstrates how language models can combine reasoning and action to solve tasks through interaction with external environments (Yao et al., 2022). Toolformer shows that language models can learn to use external tools, while Gorilla demonstrates the ability of large language models to connect with large numbers of APIs for task completion (Patil et al., 2024; Schick et al., 2023). These developments provide the technical foundation for agentic AI systems that can operate beyond simple conversational outputs.

In cybersecurity, agentic AI offers major opportunities. It can support automated threat detection, vulnerability assessment, security operations center workflows, incident triage, phishing detection, malware analysis, cloud security monitoring, and response recommendation. For example, an agentic cybersecurity system may collect alerts from multiple security tools, retrieve threat intelligence, analyze suspicious activity, prioritize incidents, recommend response actions, and escalate severe cases to human analysts. This can reduce manual workload and improve response speed in environments where security teams face large volumes of alerts and complex attack patterns.

However, the same capabilities that make agentic AI useful for cybersecurity also create new security concerns. Because agentic AI may interact with APIs, code repositories, databases, cloud platforms, security dashboards, and internal enterprise systems, it can become an active participant in cybersecurity operations. If poorly governed, such systems may execute unsafe actions, expose sensitive information, follow malicious instructions, or access tools beyond their intended permissions. The movement toward agentic AI therefore represents both an opportunity and a risk. It can strengthen cybersecurity defense, but it also expands the attack surface by introducing autonomous systems that reason, act, retrieve information, and interact with critical digital infrastructure.

1.2 Problem Statement

Existing cybersecurity controls were largely designed for human users, conventional software applications, and predictable automation systems. These controls often assume that system behavior is clearly defined, access permissions are stable, and user actions can be traced to identifiable human or software actors. Agentic AI challenges these assumptions because autonomous agents can make decisions, interact with tools, call

APIs, execute code, retrieve information, and adapt their behavior during operation. This creates a new class of cybersecurity risk that traditional security governance may not fully address.

One major concern is prompt injection, where malicious instructions are inserted into user prompts, external documents, websites, emails, or retrieved content in order to manipulate an AI system's behavior. This risk is especially serious for agentic AI because agents may retrieve and act on untrusted external information. Greshake et al. (2023) showed that indirect prompt injection can compromise real-world LLM-integrated applications by embedding malicious instructions in content processed by the model. OWASP Gen AI Security Project (2025) also identifies prompt injection as one of the most important risks in LLM applications. For cybersecurity systems, this means that an AI agent could be manipulated into ignoring security rules, leaking information, making unsafe recommendations, or using tools in unauthorized ways.

Another important risk is insecure tool use. Agentic AI systems often rely on external tools, plugins, APIs, and connected workflows. If these tools are not properly restricted, authenticated, monitored, and validated, an AI agent may perform actions outside its authorized scope. This could include accessing sensitive logs, modifying system settings, querying confidential databases, interacting with production systems, or recommending unsafe response actions. Since tool-connected language models are designed to operate across digital environments, weak permission boundaries can turn an AI assistant into an unintended security liability (Patil et al., 2024; Schick et al., 2023).

Data leakage is also a major concern. Agentic AI systems may retrieve, summarize, store, or transmit sensitive information from internal systems. Prior research has shown that large language models can expose memorized training data, creating privacy and confidentiality risks (Carlini et al., 2021). In cybersecurity contexts, the exposed data may include security logs, vulnerability reports, customer information, credentials, source code, or incident response records. These risks become more serious when AI agents are connected to enterprise systems without strong access control and monitoring.

Agentic AI also inherits broader machine learning security risks, including model extraction, poisoning attacks, and adversarial manipulation. Tramèr et al. (2016) demonstrated that machine learning models can be stolen through prediction APIs, while poisoning attacks show that attackers can manipulate training data to corrupt model behavior (Biggio et al., 2012; Jagielski et al., 2018). These risks are relevant because agentic AI systems often depend on models, data pipelines, retrieval systems, software packages, and external APIs. As a result, a weakness in any part of the agentic AI ecosystem can affect the reliability and security of the overall system.

The central problem is that traditional governance, access control, audit, and incident response models are not sufficient for AI agents that can reason, act, adapt, and interact with external systems. Organizations need a more specific defensive framework for agentic AI. Such a framework must address identity and access management, prompt and input security, secure tool governance, runtime monitoring, human oversight, auditability, explainability, and incident response. Without these controls, agentic AI may introduce serious risks despite its potential to improve cybersecurity operations.

1.3 Research Aim

The aim of this study is to examine the emerging cybersecurity risks, governance challenges, and defensive frameworks required for the safe, transparent, and accountable deployment of agentic AI systems in modern cybersecurity environments.

1.4 Research Objectives

The objectives of this study are to:

- ❖ Explain the concept of agentic AI and its relevance to cybersecurity operations.
- ❖ Identify the major cybersecurity risks associated with autonomous AI agents.
- ❖ 3. Examine governance challenges related to agentic AI decision-making, tool use, access control, and accountability.
- ❖ 4. Propose a defensive cybersecurity framework for managing agentic AI risks.
- ❖ 5. Recommend practical technical and governance controls for secure, transparent, and accountable agentic AI adoption.

1.5 Research Questions

This study is guided by the following research questions:

- ❖ What makes agentic AI different from traditional AI and generative AI in cybersecurity contexts?
- ❖ What are the major security risks introduced by agentic AI systems?
- ❖ What governance challenges arise from autonomous AI decision-making and tool use?
- ❖ How can organizations build defensive frameworks for secure agentic AI deployment?
- ❖ What technical and policy controls are needed to reduce agentic AI cyber risks?

1.6 Significance of the Study

This study is significant because agentic AI is becoming increasingly relevant to cybersecurity operations, yet its autonomy creates governance and security challenges that many existing controls may not fully address. Cybersecurity teams are under pressure to respond to complex threats, large alert volumes, ransomware campaigns, phishing attacks, software vulnerabilities, and cloud security risks. Agentic AI may help by improving automation, accelerating analysis, and supporting faster decision-making. However, these benefits can only be achieved safely if agentic AI systems are governed through strong security controls.

The study contributes to the cybersecurity field by clarifying the specific risks that arise when AI systems move from passive prediction or content generation toward autonomous action. It shows that agentic AI must be treated not only as a model or software tool, but also as an operational actor with access rights, decision pathways, tool permissions, and audit requirements. This is important because an AI agent that can retrieve information, call APIs, and interact with security systems may influence real cybersecurity outcomes.

The study is also important for governance and compliance. The NIST AI Risk Management Framework provides a structure for identifying, measuring, managing, and governing AI risks, while the NIST Generative AI Profile extends this focus to risks associated with generative AI systems (NIST, 2023, 2024). These frameworks are highly relevant to agentic AI because agentic systems often depend on generative AI capabilities, retrieval systems, tool use, and autonomous workflows. Similarly, OWASP identifies critical LLM application risks such as prompt injection, insecure output handling, training data poisoning, sensitive information disclosure, and supply-chain vulnerabilities (OWASP Gen AI Security Project, 2025). These risks are closely connected to agentic AI because agents may process untrusted inputs, rely on third-party components, interact with connected tools, and access sensitive environments.

In practical terms, this study provides value to cybersecurity managers, AI developers, security operations teams, auditors, policymakers, and researchers. It supports the development of secure-by-design agentic AI systems and encourages organizations to adopt controls such as zero trust access, least privilege permissions,

runtime monitoring, audit logging, human approval gates, and secure tool governance. The study also provides a foundation for future research on agentic AI red teaming, explainable agent behavior, AI identity governance, secure tool-use evaluation, and autonomous cyber defense.

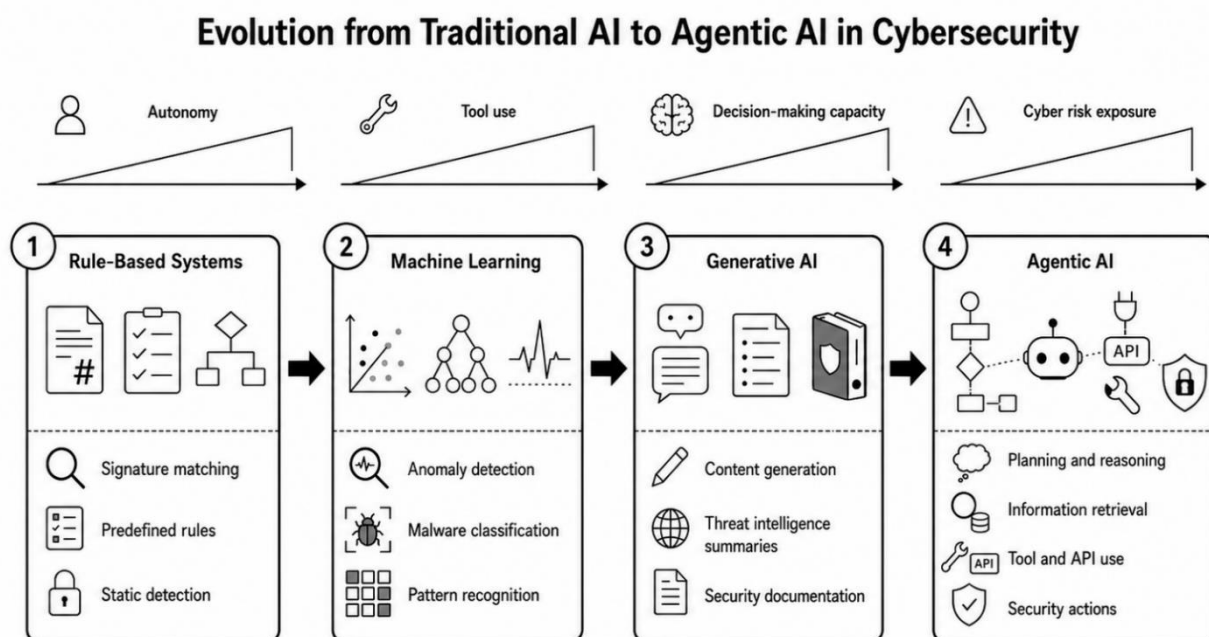


Figure 1. Evolution from traditional AI to agentic AI in cybersecurity.

The figure shows the progression from rule-based detection systems to machine learning, generative AI, and agentic AI, with increasing autonomy, tool use, operational decision-making, and cybersecurity risk exposure. Adapted conceptually from Lewis et al. (2020), Yao et al. (2022), Schick et al. (2023), and Patil et al. (2024).

2. Conceptual Background: Understanding Agentic AI in Cybersecurity

2.1 Definition of Agentic AI

Agentic artificial intelligence refers to AI systems that can pursue goals, plan actions, use tools, interact with digital environments, and adapt to feedback with limited human direction. Unlike traditional machine learning systems that mainly classify, predict, or detect patterns, agentic AI systems are designed to complete tasks through sequences of actions. They can interpret an objective, break it into steps, retrieve relevant information, select appropriate tools, execute actions, and evaluate intermediate results before deciding what to do next. Agentic AI is closely linked to recent developments in large language models and tool-connected AI systems. ReAct demonstrates that language models can combine reasoning and acting, allowing them to generate reasoning steps while interacting with external environments (Yao et al., 2022). Toolformer shows that language models can learn to use external tools to improve task performance (Schick et al., 2023). Gorilla extends this direction by connecting large language models with large numbers of APIs, making it possible for models to select and call tools for task completion (Patil et al., 2024). Together, these developments show how AI systems are moving from simple response generation toward autonomous action.

In cybersecurity, agentic AI can be understood as an AI system that can perform or assist with security-related tasks across connected digital environments. For example, an agentic AI system may receive an alert, retrieve related threat intelligence, examine system logs, compare indicators of compromise, recommend containment steps, create an incident summary, and escalate the case to a human analyst. In more advanced settings, an agent may interact with vulnerability scanners, endpoint detection systems, ticketing platforms, cloud security dashboards, or code repositories. This makes agentic AI potentially valuable for cybersecurity, but it also makes security governance more complex.

A key feature of agentic AI is that it introduces action into the AI workflow. Traditional AI may identify a suspicious pattern, and generative AI may summarize a threat report, but agentic AI may take the next step by using tools, querying systems, or recommending operational responses. This means agentic AI must be evaluated not only based on accuracy or usefulness, but also based on safety, permissions, accountability, and control. In cybersecurity contexts, an incorrect or unauthorized action can have serious consequences, including service disruption, data exposure, privilege misuse, or weakened defense posture.

2.2 Core Features of Agentic AI

Agentic AI has several core features that distinguish it from traditional AI and basic generative AI. These features include autonomy, goal-directed reasoning, tool and API use, memory, retrieval, planning, adaptive behavior, and human oversight.

First, agentic AI involves a higher level of autonomy. The system is not only responding to a single input, but may also carry out a sequence of actions to achieve a defined goal. For example, a cybersecurity agent may be asked to investigate a suspicious login event. Instead of only explaining what suspicious login means, it may examine logs, compare the login with known user behavior, check threat intelligence sources, and prepare a response recommendation.

Second, agentic AI relies on goal-directed reasoning. It can interpret an objective and determine the steps needed to complete it. This is linked to reasoning-and-acting models such as ReAct, where the model reasons through a problem while taking actions in an environment (Yao et al., 2022). In cybersecurity, this ability is useful for multi-step workflows such as incident analysis, vulnerability prioritization, and security investigation.

Third, agentic AI depends heavily on tool and API use. Tool use allows the AI system to move beyond text generation and interact with real systems. For example, an agent may call a threat intelligence API, query a security information and event management platform, search a vulnerability database, or interact with a cloud configuration tool. Toolformer and Gorilla show how language models can use tools and APIs as part of task performance (Patil et al., 2024; Schick et al., 2023). While this improves functionality, it also creates risks if tool permissions are too broad or poorly monitored.

Fourth, agentic AI may use memory and contextual learning. Memory allows an agent to retain information from previous interactions or tasks, making future responses more context-aware. In cybersecurity, memory may help agents track incident history, user behavior patterns, repeated alerts, or prior remediation actions. However, memory also creates privacy and confidentiality concerns if sensitive security data is stored or reused without adequate controls.

Fifth, retrieval is a major feature of agentic AI. Retrieval-augmented generation allows AI systems to access external knowledge sources and use retrieved content to improve their outputs (Lewis et al., 2020). In cybersecurity, retrieval can help agents access vulnerability databases, threat intelligence reports, policy

documents, incident logs, or technical documentation. However, retrieval also introduces risk because agents may process untrusted or manipulated content, making prompt injection and misinformation important concerns.

Sixth, agentic AI performs multi-step planning. Instead of responding only once, an agent may create and execute a plan. For example, it may identify the affected system, collect logs, check threat intelligence, classify the severity, recommend containment, and document the incident. Planning improves task completion, but it also requires strong oversight because errors can propagate across multiple steps.

Seventh, agentic AI can adapt in real time. It may change its actions based on new inputs, tool responses, user feedback, or environmental conditions. This makes it useful in dynamic cybersecurity environments, but it also makes behavior less predictable than fixed automation. For this reason, runtime monitoring is essential.

Finally, human-in-the-loop or human-on-the-loop supervision is a critical feature of safe agentic AI deployment. Human-in-the-loop supervision means that a human must approve certain actions before execution. Human-on-the-loop supervision means that the agent can act within defined limits, while humans monitor and intervene when necessary. In cybersecurity, high-risk actions such as changing firewall rules, disabling accounts, deleting files, escalating privileges, or modifying production systems should require human oversight.

2.3 Agentic AI Applications in Cybersecurity

Agentic AI has several potential applications in cybersecurity. One major application is automated threat detection and investigation. Security teams often deal with large numbers of alerts from endpoint detection systems, firewalls, cloud platforms, and security information and event management tools. Agentic AI can help by collecting related evidence, linking alerts to known threat patterns, summarizing incidents, and prioritizing cases for analysts. This can reduce alert fatigue and improve response speed.

Another application is vulnerability analysis and prioritization. Organizations often face thousands of vulnerabilities across software, cloud infrastructure, networks, and endpoints. Agentic AI can assist by retrieving vulnerability information, checking exploitability, comparing affected assets, reviewing business criticality, and recommending remediation priorities. This can help security teams focus on the vulnerabilities that pose the greatest practical risk.

Agentic AI can also support phishing detection and social engineering analysis. It can review suspicious emails, identify malicious indicators, compare content with known phishing patterns, extract URLs or attachments for analysis, and generate user-friendly explanations. Since phishing attacks increasingly use convincing language and personalized content, agentic AI may help improve detection and awareness when combined with proper validation.

In malware analysis, agentic AI can assist by summarizing static or dynamic analysis outputs, comparing suspicious files with known malware behaviors, and helping analysts interpret technical findings. It can also support incident response by preparing timelines, drafting containment recommendations, and generating post-incident reports. These tasks often require connecting information from multiple sources, which makes agentic AI useful when deployed safely.

Cloud security monitoring is another important area. Modern organizations use cloud services, containers, identity platforms, and distributed infrastructure. Agentic AI can help identify misconfigurations, unusual access activity, exposed resources, excessive permissions, or policy violations. However, cloud environments are highly sensitive, so agent access must be carefully restricted and monitored.

Agentic AI may also support red-team and blue-team simulations. For defensive teams, it can help simulate attacker behavior at a high level, generate training scenarios, support tabletop exercises, and improve security readiness. For blue teams, it can help test detection rules, summarize attack paths, and recommend improvements to defensive controls. These uses should remain within ethical and authorized environments. Despite these benefits, agentic AI applications in cybersecurity require strong governance. Studies on AI-assisted software development show that AI-generated outputs may introduce insecure code or unreliable vulnerability repair recommendations if users rely on them without careful validation (Pearce et al., 2023; Pearce et al., 2025; Perry et al., 2023). This is highly relevant to cybersecurity because incorrect recommendations or unsafe automated actions may weaken rather than strengthen security. Therefore, agentic AI should be used as a governed security assistant, not as an uncontrolled replacement for expert judgment.

2.4 Difference Between Traditional AI, Generative AI, and Agentic AI

Traditional AI, generative AI, and agentic AI differ in their functions, autonomy, security risks, and governance needs. Traditional AI is mainly used for prediction, classification, and pattern recognition. In cybersecurity, it supports tasks such as malware classification, spam detection, fraud detection, intrusion detection, and anomaly detection. These systems usually operate within predefined workflows and do not independently take complex actions across digital systems.

Generative AI expands AI capabilities by creating new content, such as text, code, summaries, explanations, and recommendations. In cybersecurity, generative AI can help summarize threat intelligence, draft incident reports, explain suspicious activity, assist with security awareness content, and support code review. However, generative AI can produce inaccurate, hallucinated, or insecure outputs. It can also be vulnerable to prompt injection and data leakage if not properly controlled.

Agentic AI goes further by combining reasoning, planning, retrieval, tool use, and action. In cybersecurity, agentic AI may investigate alerts, retrieve external intelligence, call APIs, interact with tools, and recommend or execute response steps. This higher level of autonomy creates greater operational value, but it also creates greater risk. Agentic AI requires runtime monitoring, access control, auditability, human oversight, and clear accountability because it can interact directly with systems and workflows.

Table 1. Comparison of Traditional AI, Generative AI, and Agentic AI in Cybersecurity

Dimension	Traditional AI	Generative AI	Agentic AI
Main function	Pattern detection and prediction	Content generation and explanation	Autonomous task execution
Cybersecurity use	Malware detection, anomaly detection, fraud detection, intrusion detection	Threat intelligence summaries, phishing analysis, incident reporting, code explanation	Automated investigation, response support, tool use, API interaction, workflow execution
Level of autonomy	Low to moderate	Moderate	High
Tool use	Limited or predefined	Usually indirect or user-mediated	Direct interaction with APIs, code, security tools, and enterprise systems
Decision process	Based on trained patterns or rules	Based on generated responses from model input	Based on goal interpretation, planning, tool use, retrieval, and feedback
Main security risk	Model error, false positives, false negatives	Hallucination, prompt injection, insecure output, data exposure	Unauthorized action, privilege escalation, unsafe automation, tool misuse, system compromise
Governance need	Model validation and performance monitoring	Output control, prompt safety, content validation	Runtime monitoring, least privilege access, human oversight, audit logging, and accountability
Example cybersecurity task	Detecting abnormal network traffic	Summarizing a phishing email	Investigating a phishing alert, checking URLs, retrieving threat intelligence, and escalating the case

Note: The comparison is informed by research on retrieval-augmented generation, reasoning-and-acting language models, tool-using language models, and API-connected foundation models (Lewis et al., 2020; Liang et al., 2024; Patil et al., 2024; Schick et al., 2023; Yao et al., 2022).

3. Emerging Cybersecurity Risks of Agentic AI

Agentic artificial intelligence introduces a new security risk environment because it does not only generate outputs, but may also reason, retrieve information, interact with tools, call APIs, and initiate actions across connected systems. In cybersecurity, this creates a dual-use problem. On one hand, agentic AI can strengthen threat detection, incident triage, vulnerability analysis, and security automation. On the other hand, the same autonomy can create risks when agents process malicious instructions, access sensitive systems, misuse tools, leak data, or perform unsafe actions. These risks are more serious than many risks associated with traditional AI because agentic AI can operate as an active system actor rather than a passive analytical tool.

3.1 Prompt Injection and Instruction Manipulation

Prompt injection is one of the most important risks facing agentic AI systems. It occurs when an attacker inserts malicious instructions into a prompt, document, website, email, code comment, data source, or retrieved content in order to manipulate the behavior of an AI system. In traditional chatbot settings, prompt injection may cause the model to produce an inappropriate or inaccurate response. In agentic AI systems, the consequences can be more serious because the agent may have access to tools, APIs, internal files, databases, or security workflows.

Prompt injection becomes especially dangerous when an AI agent retrieves or processes external information. For example, an agent may be instructed to summarize a webpage, analyze an email, review a support ticket, or retrieve threat intelligence from an external source. If that source contains hidden malicious instructions, the agent may treat them as valid commands. This is known as indirect prompt injection. Greshake et al. (2023) showed that indirect prompt injection can compromise real-world LLM-integrated applications, especially when models interact with untrusted external content. This creates a major concern for cybersecurity agents because they often need to process suspicious emails, unknown files, external threat intelligence, or user-submitted reports.

OWASP Gen AI Security Project (2025) identifies prompt injection as a major LLM application risk because it can override intended instructions, bypass safeguards, or manipulate model behavior. For agentic AI, prompt injection may lead to unauthorized tool use, disclosure of sensitive information, incorrect security recommendations, or unsafe automation. A malicious instruction may attempt to make the agent ignore security policies, reveal hidden system instructions, access restricted data, or perform actions outside its approved scope.

Instruction manipulation is also a governance issue. Agentic AI systems often rely on layered instructions, including system instructions, developer instructions, organizational policies, user requests, and retrieved content. When these instruction layers conflict, the system must be able to prioritize trusted instructions over untrusted inputs. Without strong instruction hierarchy, prompt filtering, input validation, and runtime monitoring, agentic AI may follow malicious or low-trust instructions. Therefore, prompt injection defense should be treated as a core security requirement for agentic AI deployment, not as a minor content moderation issue.

3.2 Unauthorized Tool Use and Privilege Escalation

Agentic AI systems are increasingly designed to interact with external tools, software platforms, APIs, code repositories, cloud dashboards, ticketing systems, databases, and security tools. This tool-connected design allows agents to perform useful tasks, such as retrieving threat intelligence, querying logs, opening incident tickets, checking vulnerability databases, or recommending containment actions. However, it also creates serious risks when tool access is not properly restricted.

Unauthorized tool use occurs when an AI agent accesses or uses a tool beyond its intended purpose. This may happen because of weak permission design, poor authentication, excessive privileges, unclear role boundaries, or prompt manipulation. For example, a cybersecurity agent that is only supposed to summarize alerts may be given access to modify firewall rules, disable accounts, or query sensitive databases. If the agent is compromised, misdirected, or poorly configured, it may perform actions that should have required human approval.

Privilege escalation is closely related to unauthorized tool use. If an agent has access to multiple systems, it may indirectly gain broader control than intended. This risk becomes more serious as large language models are connected to large numbers of APIs and task execution environments. Patil et al. (2024) show how large language models can connect with massive APIs, while Liang et al. (2024) describe foundation models completing tasks through large-scale API connections. These developments show why tool-connected AI systems need strict access boundaries.

Zero trust architecture is highly relevant to this challenge. Rose et al. (2020) explain that zero trust is based on continuous verification, least privilege, and the removal of implicit trust. In the context of agentic AI, this means that an AI agent should not be trusted automatically because it operates inside an organization's environment. Instead, each agent should have a defined identity, limited permissions, monitored sessions, approved tools, and revocable access. High-risk tool use should require explicit human authorization.

The risk is not only technical, but also organizational. If security teams do not define which actions agents are allowed to perform, which systems they can access, and which actions require human approval, agentic AI can become a source of uncontrolled automation. For safe deployment, organizations should apply least privilege access, tool allowlists, API scope restrictions, approval gates, logging, and continuous monitoring of every agent action.

3.3 Data Leakage and Sensitive Information Disclosure

Data leakage is another major cybersecurity risk associated with agentic AI. AI agents may retrieve, process, summarize, store, or transmit information from enterprise systems. In cybersecurity settings, this information may include security logs, threat intelligence, credentials, internal emails, vulnerability reports, incident records, source code, employee information, customer data, and compliance documents. If governance controls are weak, an agent may expose sensitive information to unauthorized users, external tools, third-party services, or insecure outputs.

Large language models have already raised concerns about memorization and sensitive data exposure. Carlini et al. (2021) showed that training data can be extracted from large language models, while Carlini et al. (2022) examined memorization across neural language models. Peng et al. (2023) also explored near-duplicate sequence search for evaluating memorization in large language models. These studies show that AI systems may retain or reproduce sensitive patterns in ways that are difficult to predict.

For agentic AI, the risk is broader because the agent may not only generate text from its model, but may also retrieve live information from connected systems. For example, an agent may access internal documents to answer a user query, retrieve logs from a security system, or summarize confidential incident data. If the user is not authorized to view that information, or if the agent does not apply proper access controls, sensitive information may be disclosed.

Sensitive information disclosure may also occur through tool outputs. An agent may call an API that returns confidential data and then include that data in a response, ticket, email, report, or external system. This creates a risk of accidental disclosure, especially when the agent is connected to multiple tools or communication channels. In cybersecurity operations, such disclosure could expose vulnerabilities, system architecture, incident details, or active security weaknesses.

Data leakage prevention for agentic AI requires strong access control, data classification, output filtering, masking of sensitive information, user authorization checks, and audit logging. Agents should only retrieve data that is necessary for the task, and outputs should be reviewed or filtered when they contain confidential

information. Organizations should also define clear policies for what data agents can access, store, summarize, and transmit.

3.4 Model and Agent Supply-Chain Risks

Agentic AI systems depend on a complex supply chain. This supply chain may include foundation models, fine-tuned models, retrieval databases, plugins, APIs, datasets, orchestration frameworks, code libraries, prompt templates, third-party tools, software packages, and cloud services. A weakness in any part of this ecosystem can affect the security of the entire agentic AI system.

OWASP Gen AI Security Project (2025) identifies supply-chain vulnerabilities as a major concern for LLM applications. For agentic AI, this concern is even stronger because agents often rely on external components to complete tasks. If a plugin, API, dataset, or package is compromised, the agent may unknowingly process malicious data, call unsafe tools, or produce unreliable outputs. This is especially important in cybersecurity, where the integrity of tools and data sources is central to effective defense.

Code-generating models introduce another supply-chain risk. Spracklen et al. (2025) show that code-generating large language models may hallucinate software packages. Package hallucination can create opportunities for dependency confusion, where an attacker creates a malicious package using a name that the model suggested but that did not previously exist. If developers or automated systems install such a package, the supply chain may be compromised.

Agentic AI also depends on retrieval systems. If the retrieval index contains outdated, manipulated, or malicious information, the agent may generate flawed analysis or follow unsafe recommendations. Similarly, if an agent is connected to third-party APIs without verification, the system may rely on inaccurate or compromised outputs. These risks show that agentic AI security must include verification of models, datasets, tools, APIs, packages, and orchestration components.

To reduce supply-chain risks, organizations should apply vendor validation, software bill of materials practices, dependency scanning, package verification, model provenance checks, tool approval processes, secure API management, and continuous monitoring. Agentic AI should not be allowed to install packages, execute code, or rely on third-party tools without defined security controls and approval procedures.

3.5 Autonomous Offensive Cyber Capabilities

Agentic AI has potential dual-use capabilities. While it can support defensive cybersecurity operations, it may also be misused to automate harmful cyber activities. At a high level, agentic AI could support reconnaissance, phishing preparation, vulnerability discovery, social engineering analysis, malware-related assistance, or attack planning. This does not mean that agentic AI should be treated only as a threat, but it does mean that organizations must recognize how autonomy can increase misuse potential.

Existing research on machine learning security shows that AI systems can be attacked, manipulated, extracted, or used in adversarial ways. Tramèr et al. (2016) demonstrated that models can be stolen through prediction APIs. Papernot et al. (2017) examined practical black-box attacks against machine learning systems, while Papernot et al. (2018) reviewed broader security and privacy concerns in machine learning. Biggio et al. (2012) also showed how poisoning attacks can manipulate support vector machines. These studies show that AI systems are not only tools for defense, but also targets and possible enablers of cyber risk.

The concern with agentic AI is that it may reduce the effort required to perform multi-step cyber tasks. A malicious actor may attempt to use an agent to gather information, organize attack steps, generate social

engineering content, or analyze weaknesses. While ethical restrictions and safety filters may reduce this risk, determined attackers may attempt to bypass controls through prompt manipulation, model misuse, or unauthorized tool integration.

For this reason, research and governance must focus on safe boundaries for agentic AI. Defensive use should be clearly separated from offensive misuse. Organizations should prohibit unauthorized exploitation, restrict high-risk tool use, monitor agent behavior, and maintain human oversight for sensitive security tasks. Agentic AI systems used in cybersecurity should be designed to support lawful, ethical, and authorized defensive operations only.

3.6 Hallucinated or Unsafe Security Actions

Agentic AI systems may produce hallucinated, inaccurate, or unsafe outputs. In cybersecurity, this is particularly dangerous because incorrect recommendations may affect system availability, user access, incident response, or organizational security posture. An agent may falsely classify legitimate activity as malicious, recommend blocking a safe domain, suggest deleting important files, misinterpret a vulnerability, or propose a configuration change that creates new risk.

This risk becomes more serious when the agent can act on its own recommendations. A generative AI system may produce an incorrect answer, but an agentic AI system may use tools or trigger workflows based on that answer. If the agent has permission to modify systems, disable accounts, open tickets, update policies, or initiate containment actions, hallucinated reasoning can become operational damage.

AI-assisted coding studies are relevant here. Pearce et al. (2025) assessed the security of GitHub Copilot's code contributions and highlighted concerns about insecure code generation. Pearce et al. (2023) examined the use of large language models for vulnerability repair, showing that automated repair is not always reliable. Perry et al. (2023) found that users may write more insecure code when using AI assistants. These findings show that AI outputs should not be accepted automatically in security-sensitive contexts.

Unsafe security actions may also arise from incomplete context. An agent may recommend action based on limited evidence, outdated threat intelligence, or incomplete logs. For example, it may interpret unusual login activity as malicious without considering legitimate travel, maintenance work, or business context. This is why human oversight, validation checks, confidence scoring, and approval gates are necessary for high-risk actions.

Organizations should treat agentic AI recommendations as decision support, not final authority. Agents should be allowed to assist with analysis, summarization, prioritization, and documentation, but destructive or high-impact actions should require human review. Safe deployment requires validation layers, rollback procedures, incident simulation, and continuous performance monitoring.

3.7 Lack of Accountability and Explainability

Accountability is a major challenge in agentic AI because the system may perform multiple actions across several tools and decision points. When an AI agent retrieves data, reasons through a task, calls APIs, recommends action, and generates a report, it may be difficult to determine why a decision was made, which information influenced the output, and who approved the final action. This creates serious concerns for audit, compliance, incident response, and legal defensibility.

Raji et al. (2020) emphasize the importance of internal algorithmic auditing for closing accountability gaps. Mitchell et al. (2019) propose model cards as a way to improve model reporting and transparency. Mökander

et al. (2022) also highlight the value of conformity assessments and post-market monitoring in AI governance. These approaches are important for agentic AI because organizations need clear documentation of model behavior, system limits, risks, and post-deployment performance.

Explainability is especially important in cybersecurity because analysts must understand why an agent made a recommendation. If an agent recommends blocking an IP address, disabling an account, escalating an incident, or applying a patch, the security team needs to know the evidence behind that decision. Without explainability, analysts may either trust the system blindly or reject it entirely. Both outcomes are problematic. Weak auditability also makes incident response difficult. If an agent causes an error, leaks data, or performs an unauthorized action, investigators need logs showing prompts, retrieved data, tool calls, outputs, approvals, and final actions. Without these records, it may be impossible to determine whether the problem came from prompt injection, model error, user misuse, faulty data, tool failure, or weak governance.

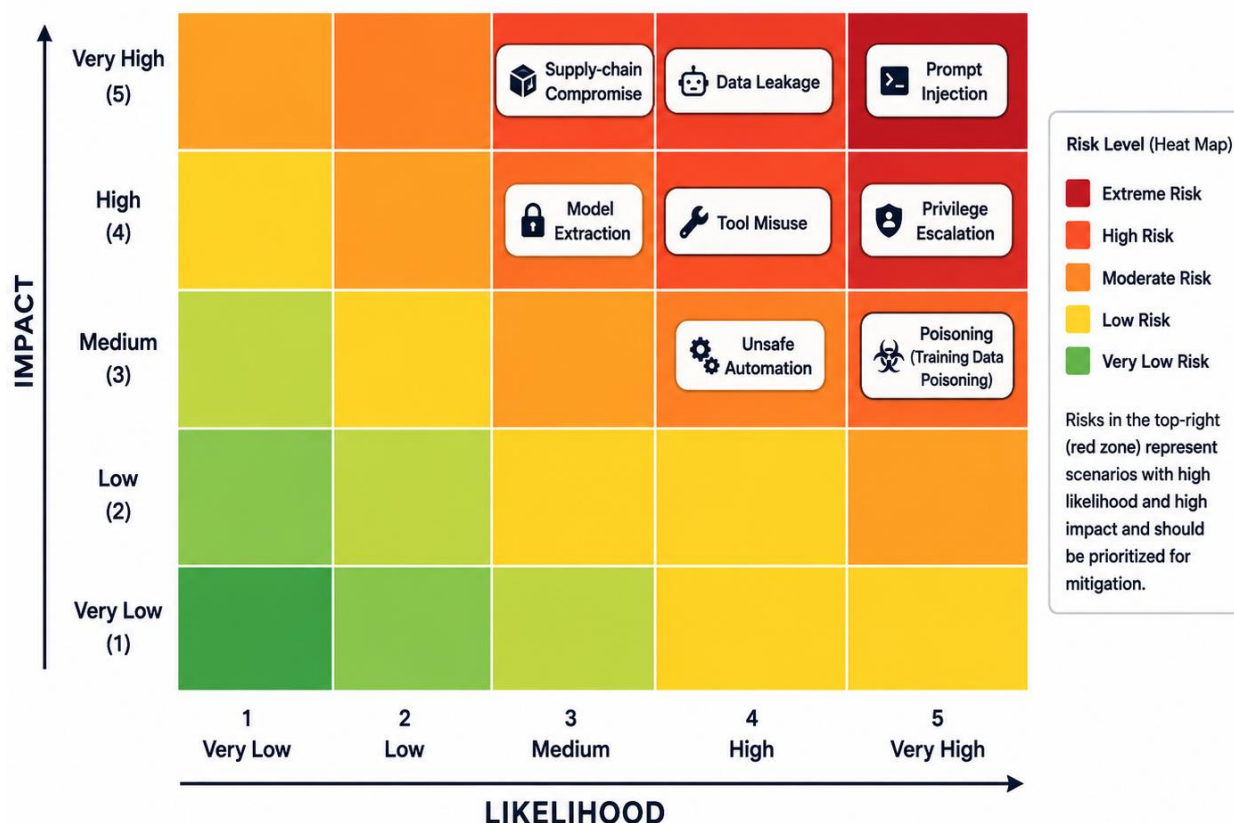
To address this risk, organizations should maintain detailed agent activity logs, decision records, tool-call histories, access records, human approval records, and incident review procedures. Explainability should not be treated as an optional feature. It is a core requirement for trustworthy agentic AI in cybersecurity.

Table 2. Emerging Risks of Agentic AI in Cybersecurity

Risk Category	Description	Cybersecurity Impact	Example Control
Prompt injection	Malicious instruction manipulation through direct or indirect prompts	Unauthorized behavior, policy bypass, unsafe outputs	Prompt filtering and instruction hierarchy
Tool misuse	Unsafe or unauthorized API, plugin, or system action	Privilege abuse, unauthorized system change, operational disruption	Least privilege access and tool allowlists
Data leakage	Exposure of sensitive data through retrieval, memory, or outputs	Privacy violation, compliance risk, disclosure of confidential security data	Data loss prevention and output filtering
Supply-chain risk	Compromised model, plugin, package, dataset, or third-party tool	System compromise, unreliable outputs, malicious dependency exposure	Vendor validation and software bill of materials
Unsafe automation	Incorrect classification, recommendation, or response action	Business disruption, misconfiguration, false containment	Human approval gates and rollback procedures
Weak auditability	Poor traceability of agent decisions, tool calls, and approvals	Accountability failure, weak incident investigation, compliance gaps	Agent activity logging and decision records

Note: Risk categories are informed by OWASP Gen AI Security Project (2025), Greshake et al. (2023), Carlini et al. (2021), Spracklen et al. (2025), and Raji et al. (2020).

Risk Severity Matrix for Agentic AI Cybersecurity Threats



Based on risk categories from NIST (2023, 2024), OWASP Gen AI Security Project (2025), and ENISA (2025).

The chart maps key agentic AI risks according to likelihood and impact, including prompt injection, tool misuse, data leakage, supply-chain compromise, unsafe automation, poisoning, model extraction, and privilege escalation. Based on risk categories from NIST (2023, 2024), OWASP Gen AI Security Project (2025), and ENISA (2025).

4. Governance Challenges in Agentic AI Cybersecurity

Governance is central to the safe use of agentic AI in cybersecurity because technical controls alone cannot fully manage the risks of autonomous decision-making, tool use, data access, and operational action. Agentic AI systems require clear policies, defined responsibilities, access boundaries, monitoring mechanisms, audit procedures, and human oversight. Without governance, organizations may deploy AI agents that are powerful but poorly controlled. This can create accountability gaps, compliance uncertainty, and operational risk.

4.1 Accountability for Autonomous Decisions

A major governance challenge is determining who is responsible when an AI agent makes a decision or performs an action. In traditional cybersecurity workflows, responsibility is often easier to assign because human analysts, administrators, or software systems operate within defined roles. With agentic AI, responsibility may be distributed across model developers, system integrators, security teams, vendors, data owners, compliance officers, and organizational leadership.

For example, if an AI agent recommends blocking a legitimate user account, who is accountable for the decision? The developer who built the agent, the security team that deployed it, the vendor that supplied the model, or the manager who approved its use? The answer is not always clear. This becomes more complex when the agent uses multiple tools, retrieves external information, or adapts its response based on changing inputs.

Raji et al. (2020) argue that internal algorithmic auditing is needed to close accountability gaps in AI systems. This is directly relevant to agentic AI because organizations must be able to examine how an agent was designed, tested, deployed, monitored, and reviewed. Mitchell et al. (2019) also propose model cards as a transparency mechanism for documenting model purpose, performance, limitations, and appropriate use. For agentic AI, similar documentation should include not only model details, but also tool permissions, action boundaries, escalation rules, and monitoring requirements.

Accountability requires clear ownership. Organizations should define who is responsible for agent configuration, access approval, incident review, tool integration, data governance, and performance monitoring. They should also define who has authority to approve high-risk agent actions. Without this structure, autonomous AI systems may create a gap between technical capability and organizational responsibility.

4.2 Runtime Governance and Continuous Monitoring

Pre-deployment testing is necessary, but it is not enough for agentic AI. Traditional software can often be tested against expected inputs and outputs before release. Agentic AI systems are more dynamic because they may process new prompts, retrieve changing information, interact with tools, and adapt to different contexts during operation. This means that safe behavior during testing does not guarantee safe behavior after deployment.

Sculley et al. (2015) describe hidden technical debt in machine learning systems, including risks related to data dependencies, configuration complexity, monitoring gaps, and system maintenance. Kreuzberger et al. (2023) explain that MLOps provides practices for deployment, monitoring, lifecycle management, and operational control of machine learning systems. These insights are important for agentic AI because agents operate across models, tools, data pipelines, APIs, and organizational workflows.

Runtime governance means that agentic AI systems should be monitored continuously after deployment. Organizations need to observe what agents are doing, what tools they are calling, what data they are accessing, what instructions they are following, and whether their behavior deviates from approved boundaries. Continuous monitoring can detect prompt injection attempts, unusual tool use, repeated failed actions, excessive data access, abnormal response patterns, or unauthorized workflow execution.

Runtime governance should also include drift detection. Agent behavior may change because of updated models, new data sources, changed APIs, modified prompts, altered permissions, or changing threat environments. If these changes are not monitored, the organization may not notice when the agent begins producing unreliable outputs or taking unsafe actions. Therefore, agentic AI requires continuous validation, not one-time approval.

Effective runtime governance includes logging, alerting, anomaly detection, human review, periodic audits, and incident response procedures. Agents should have defined operational boundaries, and violations should trigger review or automatic containment. In high-risk environments, organizations should also use sandboxing and staged deployment before allowing agents to operate in production systems.

4.3 Identity and Access Management for AI Agents

Identity and access management is one of the most important governance issues for agentic AI. Since AI agents may access tools, databases, cloud platforms, code repositories, and security systems, they should be treated as digital identities. They should not share generic service accounts, administrator credentials, or broad access tokens. Each agent should have a unique identity, defined permissions, authentication requirements, access boundaries, and revocation procedures.

Zero trust architecture provides a useful foundation for this approach. Rose et al. (2020) explain that zero trust requires continuous verification, least privilege access, and segmentation. In agentic AI governance, this means that agents should only access the systems and data needed for their approved tasks. Access should be limited by role, context, risk level, and task type. The agent should not receive broad permissions simply because it is part of an internal system.

This is especially important in cybersecurity because agents may operate across sensitive environments. An agent connected to a security dashboard may see threat alerts, network logs, user activity, device information, and incident records. An agent connected to a code repository may access proprietary software or security-related source code. An agent connected to a cloud platform may view configurations, permissions, storage resources, or workload details. If these permissions are excessive, the agent becomes a high-value target.

Access management should include authentication, authorization, session control, credential protection, tool scoping, and access review. Organizations should also maintain logs of every tool call and data access event. When an agent no longer requires access, its permissions should be revoked immediately. If an agent behaves unexpectedly, security teams should be able to disable its access quickly.

AI agents should also be separated by function. A monitoring agent should not automatically have the same privileges as a remediation agent. A reporting agent should not be able to modify production systems. This separation supports least privilege and reduces the impact of errors or compromise.

4.4 Compliance and Regulatory Alignment

Agentic AI governance must align with existing cybersecurity, privacy, and AI risk management requirements. Organizations cannot deploy autonomous AI agents without considering legal, regulatory, and policy obligations. This is especially important when agents process personal data, sensitive business information, regulated records, or security-critical infrastructure.

NIST's AI Risk Management Framework provides a general structure for identifying, measuring, managing, and governing AI risks (NIST, 2023). The NIST Generative AI Profile extends this guidance by focusing on risks associated with generative AI systems, including content risks, misuse, data concerns, and system vulnerabilities (NIST, 2024). These frameworks are highly relevant because agentic AI often combines generative AI, retrieval, tool use, and autonomous workflow execution.

Zero trust architecture also supports compliance by providing a security model based on continuous verification and restricted access (Rose et al., 2020). For agentic AI, zero trust can help ensure that agents do not receive excessive permissions and that all actions are monitored. This is especially important for organizations that must comply with data protection, cybersecurity, financial, healthcare, or public sector requirements.

Mökander et al. (2022) emphasize the importance of conformity assessments and post-market monitoring in AI regulation. This is relevant because agentic AI systems may behave differently after deployment than they did during testing. Compliance therefore requires ongoing monitoring, documentation, and review.

Organizations should not treat approval as a one-time event. Instead, they should periodically assess whether the agent still performs safely, securely, and within its intended scope.

Regulatory alignment also requires documentation. Organizations should document agent purpose, data sources, tool access, risk assessments, approval rules, monitoring procedures, incident response plans, and accountability structures. This documentation can support audits, compliance reviews, and legal defensibility if an AI-related incident occurs.

4.5 Transparency and Explainability

Transparency and explainability are essential for governing agentic AI in cybersecurity. Security teams need to understand what an agent did, why it did it, what information it used, and which tools it called. Without transparency, it becomes difficult to trust, audit, correct, or defend agentic AI decisions.

Transparency includes clear documentation of the agent's purpose, operating boundaries, data access, tool permissions, and limitations. Explainability focuses on the ability to interpret the agent's reasoning or decision process. In cybersecurity, explainability is especially important because decisions may affect system availability, user access, incident severity, vulnerability prioritization, or containment actions.

Mitchell et al. (2019) propose model cards as a method for improving transparency in model reporting. Raji et al. (2020) emphasize internal algorithmic auditing as a way to improve accountability. Mökander et al. (2022) show that post-market monitoring and conformity assessments are important for AI governance. These approaches can support agentic AI governance by helping organizations document model behavior, review system performance, and identify risks after deployment.

For agentic AI, transparency should go beyond model-level documentation. Organizations should also maintain records of prompts, retrieved information, tool calls, action sequences, approvals, and outputs. If an agent recommends a security action, analysts should be able to see the evidence behind the recommendation. If an agent accesses sensitive data, auditors should be able to review when, why, and under which authorization it did so.

Explainability also improves human oversight. Analysts are more likely to trust agentic AI when they can review the reasoning behind its output. However, explainability should not be confused with blind acceptance. Even if an agent provides a clear explanation, its recommendation should still be validated when the action is high risk.

4.6 Human Oversight and Approval Mechanisms

Human oversight is necessary because agentic AI systems may make mistakes, misinterpret information, follow malicious prompts, or recommend actions that carry operational risk. In cybersecurity, some actions should never be fully automated without approval. These include actions involving sensitive systems, user account changes, financial systems, production infrastructure, patch deployment, incident containment, firewall modification, data deletion, or security policy changes.

NIST's AI risk management guidance emphasizes governance, measurement, management, and oversight as part of reducing harmful outcomes (NIST, 2023, 2024). For agentic AI, this means that organizations should define which actions can be automated, which actions require human review, and which actions are prohibited. Human approval gates are especially important for high-impact decisions.

There are two common oversight models. Human-in-the-loop oversight means that a human must approve an action before the AI agent executes it. Human-on-the-loop oversight means that the agent can operate within

defined boundaries while humans monitor activity and intervene when necessary. Both models can be useful, but the choice depends on the risk level of the task. Low-risk tasks, such as summarizing alerts or drafting incident reports, may require lighter oversight. High-risk tasks, such as disabling accounts or changing cloud configurations, should require explicit human approval.

Human oversight should also include escalation procedures. If an agent detects uncertainty, conflicting evidence, high-risk behavior, or potential policy violation, it should escalate to a human analyst. The system should also allow humans to pause, override, or disable the agent when necessary.

Effective oversight depends on clear policies and usable interfaces. Analysts should be able to review the agent's evidence, reasoning, confidence level, proposed action, and potential impact before approving action. This ensures that agentic AI supports human judgment rather than replacing it in situations where accountability and context are essential.

Table 3. Governance Challenges and Recommended Controls for Agentic AI

Governance Challenge	Why It Matters	Recommended Control
Accountability gap	It may be difficult to assign responsibility for autonomous agent decisions and actions	Clear ownership, approval policy, and internal algorithmic auditing
Runtime unpredictability	Agent behavior may change during operation due to new inputs, tools, data, or contexts	Continuous monitoring, drift detection, and behavior analytics
Excessive privileges	Agents may access more systems, tools, or data than needed for their tasks	Least privilege access, role-based access control, and tool scoping
Poor explainability	Decisions may be difficult to audit, validate, or defend	Explainable logging, decision records, and model documentation
Weak human oversight	Unsafe actions may occur automatically without expert review	Human-in-the-loop approval and escalation procedures
Compliance uncertainty	Existing rules may not directly address autonomous AI agents	AI risk management framework, conformity assessment, and post-deployment monitoring

Note: Governance controls are supported by NIST (2023, 2024), Rose et al. (2020), Raji et al. (2020), Mitchell et al. (2019), and Mökander et al. (2022).

Agentic AI Governance Model for Cybersecurity

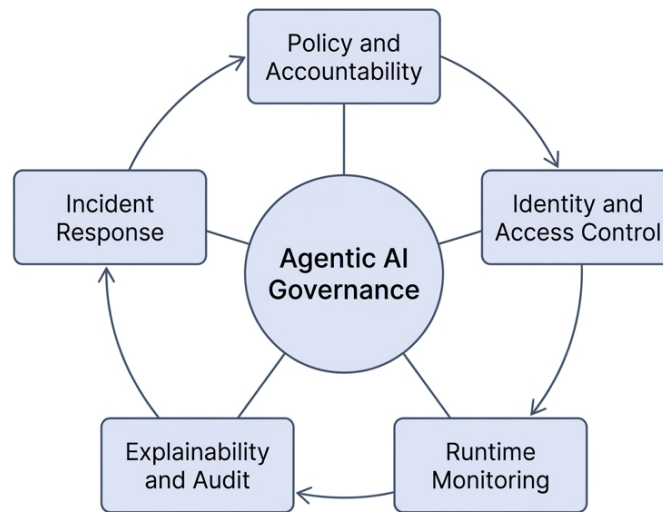


Figure 2: Agentic AI Governance Model for Cybersecurity

Figure 2. Agentic AI governance model for cybersecurity, showing five connected governance components: policy and accountability, identity and access control, runtime monitoring, explainability and audit, and incident response.

5. Defensive Framework for Secure Agentic AI Deployment

The increasing autonomy of agentic artificial intelligence systems creates a need for stronger defensive architectures that combine cybersecurity controls, AI governance, runtime monitoring, explainability, and human oversight. Unlike traditional AI systems that mainly generate outputs or perform narrow analytical tasks, agentic AI systems can interact with tools, retrieve information, access APIs, execute workflows, and make operational decisions. As a result, organizations require layered security frameworks capable of reducing risks related to prompt injection, unauthorized tool use, privilege escalation, unsafe automation, and data leakage.

A secure deployment strategy for agentic AI should not rely on a single control mechanism. Instead, it should combine governance, technical safeguards, continuous monitoring, and incident response across the entire AI lifecycle. Existing research on AI risk management, zero trust architecture, MLOps governance, and LLM application security provides useful foundations for developing secure and accountable agentic AI systems (Kreuzberger et al., 2023; NIST, 2023, 2024; OWASP Gen AI Security Project, 2025; Rose et al., 2020).

5.1 Framework Overview

The proposed defensive framework is designed as a layered governance and security architecture for agentic AI deployment in cybersecurity environments. The framework combines six interconnected layers: agent identity and access control, secure tool and API governance, prompt and input security, runtime monitoring and behavioral analytics, human oversight and escalation, and audit with incident response. These layers work together to reduce operational risks while maintaining accountability and continuous monitoring.

The framework is grounded in four major principles. First, AI risk management provides structured processes for identifying, measuring, managing, and governing AI-related risks throughout the deployment lifecycle (NIST, 2023, 2024). Second, zero trust architecture emphasizes least privilege access, continuous verification, and restricted trust boundaries (Rose et al., 2020). Third, LLM application security highlights emerging risks such as prompt injection, insecure outputs, and supply-chain vulnerabilities (OWASP Gen AI Security Project, 2025). Fourth, MLOps governance supports lifecycle monitoring, deployment management, and continuous operational oversight for machine learning systems (Kreuzberger et al., 2023).

The framework assumes that agentic AI systems should be treated as operational actors within cybersecurity environments rather than passive analytical tools. This means that agents require identities, permissions, monitoring, logging, and oversight comparable to those applied to privileged human users or enterprise software systems. The framework also recognizes that governance is continuous rather than static. Agent behavior may evolve because of changing inputs, retrieval systems, APIs, tools, or environmental conditions. Therefore, secure deployment requires continuous monitoring and adaptive governance.

5.2 Layer 1: Agent Identity and Access Control

The first layer of the framework focuses on identity and access control. Every agentic AI system should have a unique digital identity with clearly defined permissions, access limits, session boundaries, and revocation mechanisms. AI agents should never operate through shared credentials or unrestricted administrator privileges because excessive access increases the risk of misuse, compromise, or unintended actions.

This layer aligns closely with zero trust architecture principles described by Rose et al. (2020). Zero trust assumes that no user, system, or device should be trusted automatically. Instead, access decisions should be based on continuous verification, least privilege, and contextual authorization. In the context of agentic AI, this means that agents should only access the tools, systems, APIs, and data necessary for their approved tasks. Identity management for AI agents should include authentication, authorization, access scoping, session monitoring, and access revocation. Organizations should maintain clear records of which systems an agent can access, which APIs it can call, and which actions it is permitted to perform. Permissions should also be separated according to operational roles. For example, a monitoring agent should not automatically have the same permissions as a remediation or deployment agent.

Strong identity governance reduces the impact of prompt injection, unauthorized tool use, and privilege escalation. If an agent becomes compromised or behaves unexpectedly, restricted permissions can prevent the system from performing destructive or unauthorized actions. Organizations should also implement short-lived credentials, segmented environments, and continuous access reviews to reduce long-term exposure.

5.3 Layer 2: Secure Tool and API Governance

The second layer focuses on governance of tools, APIs, plugins, and connected services. Agentic AI systems increasingly depend on external tools for retrieving information, executing workflows, accessing databases, interacting with cloud services, or generating operational outputs. While this capability improves functionality, it also increases security exposure because the agent can interact directly with sensitive systems. Liang et al. (2024), Patil et al. (2024), and Schick et al. (2023) demonstrate how tool-connected and API-enabled language models can complete tasks across multiple systems. However, these capabilities create risks when agents are granted unrestricted access or when connected tools are not validated properly.

To reduce these risks, organizations should implement approved tool lists, API validation, role-based tool permissions, and execution restrictions. Agents should only access trusted tools that have undergone security review and risk assessment. High-risk APIs should require additional authorization, while sensitive actions should trigger approval workflows before execution.

Secure tool governance should also include logging and validation. Every tool call should be recorded with details such as timestamp, agent identity, accessed resource, requested action, and execution result. Validation mechanisms should confirm that outputs returned by external tools are trustworthy and compliant with organizational policies.

API segmentation is equally important. Agents should not receive unrestricted access to broad enterprise APIs because a compromised or manipulated agent could perform harmful actions across multiple systems. Instead, APIs should expose limited, task-specific functionality with restricted permissions and monitoring.

5.4 Layer 3: Prompt and Input Security

Prompt and input security forms the third layer of the defensive framework. Since agentic AI systems rely heavily on prompts, retrieved content, external instructions, and user interactions, they are vulnerable to prompt injection and instruction manipulation attacks.

Greshake et al. (2023) show how both direct and indirect prompt injection can compromise LLM-integrated applications. OWASP Gen AI Security Project (2025) also identifies prompt injection as one of the most critical risks affecting LLM applications. For agentic AI, the risk becomes more severe because the system may not only generate outputs but also perform operational actions based on manipulated instructions.

Prompt and input security should therefore include multiple defensive mechanisms. First, systems should implement prompt filtering and sanitization to detect malicious instructions, hidden commands, and unsafe requests. Second, instruction hierarchy should ensure that trusted system instructions override external or user-provided content. Third, retrieval systems should validate and classify external information before it is processed by the agent.

Organizations should also isolate high-risk content sources. External documents, emails, websites, or retrieved information should not automatically be treated as trusted input. Agents interacting with untrusted content should operate within restricted environments and limited permission scopes.

Additional controls may include adversarial testing, content validation, anomaly detection, and runtime policy enforcement. These controls help reduce the risk of manipulated prompts causing unauthorized tool use, data leakage, or unsafe operational decisions.

5.5 Layer 4: Runtime Monitoring and Behavioral Analytics

The fourth layer focuses on runtime monitoring and behavioral analytics. Pre-deployment testing alone cannot guarantee safe operation because agentic AI systems may adapt to changing contexts, interact with new data sources, and behave differently after deployment.

Kreuzberger et al. (2023) emphasize the importance of MLOps lifecycle management for machine learning systems, while Sculley et al. (2015) describe hidden technical debt associated with evolving AI systems. These concerns are especially relevant for agentic AI because agents may continuously interact with dynamic environments.

Runtime monitoring should therefore track agent behavior continuously. Organizations should monitor tool usage, API activity, data access patterns, prompt history, policy violations, behavioral drift, and unusual execution sequences. Monitoring systems should also detect indicators of compromise, such as repeated failed actions, unauthorized access attempts, suspicious outputs, or unexpected privilege usage.

Behavioral analytics can support anomaly detection by comparing current behavior against approved operational baselines. If an agent begins accessing unfamiliar systems, generating abnormal outputs, or exceeding expected task boundaries, monitoring systems should trigger alerts or containment procedures.

Drift detection is another important capability. Agent behavior may change because of model updates, new retrieval data, altered prompts, or evolving workflows. Continuous monitoring helps organizations identify when an agent deviates from approved behavior and requires investigation or retraining.

Runtime governance should also integrate with incident response systems so that suspicious activity can trigger automated containment, session termination, or escalation to human analysts.

5.6 Layer 5: Human Oversight and Escalation

Human oversight remains essential even when agentic AI systems operate autonomously. High-risk actions should require explicit human review before execution because AI systems may generate inaccurate outputs, misinterpret context, or respond unpredictably to malicious inputs.

NIST (2023, 2024) emphasizes governance structures that support accountability, oversight, and risk reduction in AI systems. Raji et al. (2020) similarly highlight the importance of governance mechanisms that reduce accountability gaps in automated systems.

Within the proposed framework, human oversight is applied through escalation thresholds and approval gates. Low-risk activities such as alert summarization or log classification may operate with limited supervision, while high-risk actions should require direct approval. Examples include deleting files, changing firewall rules, escalating privileges, deploying patches, modifying security policies, or interacting with production infrastructure.

Human oversight can operate through either human-in-the-loop or human-on-the-loop models. In human-in-the-loop systems, an agent must obtain approval before executing sensitive actions. In human-on-the-loop systems, the agent operates autonomously within approved limits while human analysts monitor its behavior and intervene when necessary.

Escalation mechanisms are equally important. If an agent encounters uncertainty, conflicting evidence, policy violations, or unexpected system responses, the system should automatically escalate the situation to human analysts. Oversight interfaces should provide sufficient explainability so that reviewers can understand why the agent proposed a specific action.

Human oversight does not eliminate automation. Instead, it creates a balance between operational efficiency and governance control.

5.7 Layer 6: Audit, Explainability, and Incident Response

The final layer focuses on auditability, explainability, and incident response. Every action performed by an agentic AI system should be traceable and reviewable. This is essential for accountability, compliance, forensic investigation, and operational trust.

Logs should capture prompts, retrieved information, accessed data, tool calls, generated outputs, approvals, policy decisions, and final actions. These records allow organizations to reconstruct agent behavior during investigations or security incidents.

Mitchell et al. (2019) emphasize the importance of model documentation and transparency through model cards. Raji et al. (2020) highlight internal algorithmic auditing, while Mökander et al. (2022) discuss post-market monitoring and conformity assessment for AI governance. These concepts provide useful guidance for maintaining accountability in agentic AI systems.

Explainability mechanisms should help analysts understand how the agent reached a decision. This may include showing retrieved evidence, reasoning chains, policy references, confidence scores, or tool usage history. Explainability improves trust and supports human oversight because analysts can validate recommendations before approval.

Incident response integration is also necessary. Organizations should establish procedures for containing compromised agents, revoking permissions, isolating affected systems, rolling back unsafe actions, and conducting post-incident reviews. Runtime monitoring systems should connect directly with incident response workflows so that suspicious activity can trigger containment automatically.

Together, auditability, explainability, and incident response create the accountability foundation required for trustworthy deployment of agentic AI in cybersecurity environments.

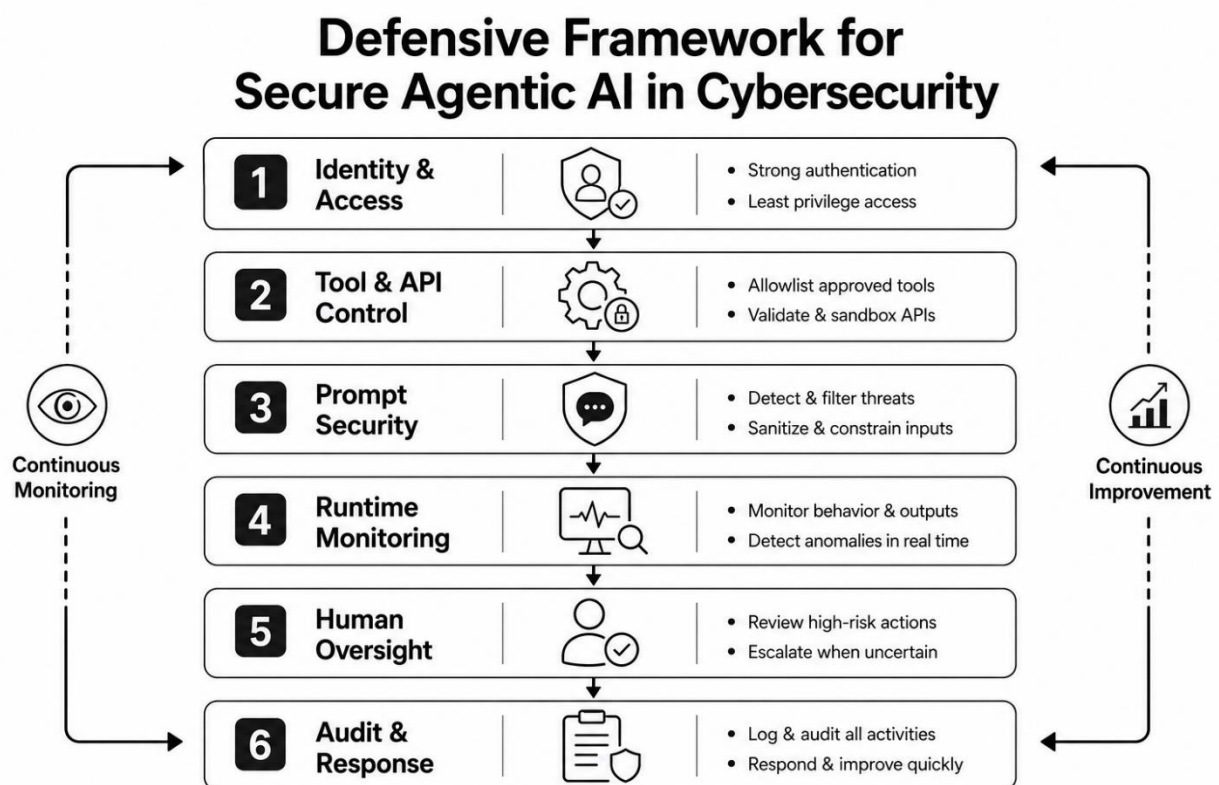


Figure : Defensive Framework for Secure Agentic AI in Cybersecurity

The framework combines identity governance, secure tool management, prompt security, runtime monitoring, human oversight, and audit mechanisms for reducing cybersecurity risks in autonomous AI systems. Framework developed from NIST (2023, 2024), OWASP Gen AI Security Project (2025), Rose et al. (2020), Kreuzberger et al. (2023), and Raji et al. (2020).

6. Practical Implementation Strategy

6.1 Risk Assessment Before Deployment

Before deploying agentic AI systems, organizations should conduct structured risk assessments that classify systems according to autonomy level, tool access, system connectivity, data sensitivity, and potential operational impact. AI systems that can interact with production infrastructure or sensitive enterprise systems should be treated as high-risk deployments requiring stronger governance controls.

NIST AI RMF provides a useful framework for identifying, measuring, managing, and governing AI risks before and after deployment (NIST, 2023, 2024). Organizations should assess risks related to data leakage, unauthorized tool use, prompt injection, operational disruption, and compliance exposure.

Risk assessment should also consider the operational context of the agent. For example, an internal reporting assistant presents different risks than an autonomous remediation agent connected to production systems. Risk classification should therefore guide access control, monitoring requirements, oversight levels, and incident response planning.

6.2 Secure Design Principles

Secure deployment of agentic AI requires secure-by-design principles integrated throughout the system lifecycle. Key principles include least privilege access, zero trust architecture, segmented environments, sandboxed execution, human approval for high-risk actions, continuous logging, and validated tool access.

Organizations should also conduct adversarial testing and red-team exercises to identify weaknesses before deployment. Research on adversarial machine learning demonstrates that AI systems are vulnerable to poisoning attacks, evasion attacks, extraction attacks, and privacy risks (Biggio et al., 2012; Carlini & Wagner, 2017; Papernot et al., 2017; Shokri et al., 2017; Tramèr et al., 2016). These findings show why AI systems must be tested against hostile conditions rather than only normal operational scenarios.

Secure design also includes supply-chain security. Organizations should validate models, datasets, APIs, plugins, and software dependencies before integration. Runtime restrictions should prevent agents from installing unauthorized packages or interacting with unapproved services.

Segmented execution environments are another important principle. High-risk agents should operate within isolated environments where unsafe actions can be contained without affecting production systems.

6.3 Monitoring Metrics for Agentic AI Security

Effective governance requires measurable security metrics that evaluate operational behavior and risk exposure after deployment. Monitoring metrics help organizations identify misuse, detect anomalies, measure governance effectiveness, and improve incident response.

Important metrics include unauthorized tool-call rates, prompt injection detection rates, abnormal behavior frequency, human escalation frequency, drift scores, data exposure incidents, and mean time to contain agent-related incidents. These metrics provide visibility into both technical performance and governance effectiveness.

Continuous monitoring is necessary because AI systems may degrade over time due to changing environments, technical debt, data cascades, or evolving operational conditions (Kreuzberger et al., 2023; Sambasivan et al., 2021; Sculley et al., 2015).

Table 4. Security Metrics for Monitoring Agentic AI Systems

Metric	Purpose	Example Measurement
Unauthorized tool-call rate	Detect misuse of connected tools	Number of blocked tool calls per week
Prompt injection detection rate	Measure input security	Percentage of malicious prompts detected
Human escalation frequency	Track high-risk agent decisions	Number of actions requiring approval
Agent drift score	Detect behavior deviation	Difference from approved behavior baseline
Sensitive data exposure events	Monitor data leakage	Number of blocked disclosure attempts
Mean time to contain agent incident	Measure response speed	Average containment time

Note: Monitoring metrics are informed by NIST (2023, 2024), OWASP Gen AI Security Project (2025), Kreuzberger et al. (2023), and Sculley et al. (2015).

6.4 Maturity Model for Agentic AI Cybersecurity

Organizations are likely to adopt agentic AI governance gradually rather than achieving mature governance immediately. A maturity model helps organizations evaluate current capabilities and identify areas requiring improvement.

The proposed maturity model contains five levels:

- ❖ **Initial:** Organizations have no formal controls for agentic AI systems. AI use is informal, poorly monitored, and weakly governed.
- ❖ **Developing:** Basic policies, limited access controls, and simple approval procedures are introduced. Monitoring capabilities remain limited.
- ❖ **Defined:** Organizations establish structured governance frameworks, monitoring procedures, role-based access control, and documented oversight policies.
- ❖ **Managed:** Continuous risk assessment, runtime monitoring, auditing, and incident response procedures are integrated into operations.
- ❖ **Optimized:** Organizations implement adaptive governance, automated monitoring, explainability mechanisms, continuous validation, and mature defensive controls across the AI lifecycle.

This maturity progression aligns with principles from AI risk management, zero trust architecture, MLOps governance, and algorithmic auditing (Kreuzberger et al., 2023; NIST, 2023; Raji et al., 2020; Rose et al., 2020).

Agentic AI Cybersecurity Governance Maturity Model

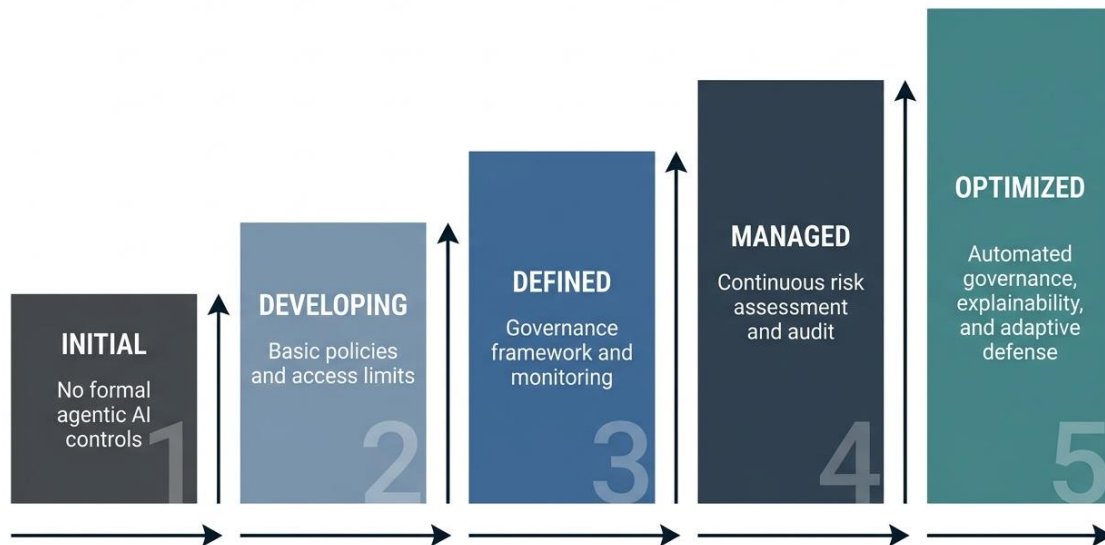


Chart 2. Agentic AI cybersecurity governance maturity model. The model illustrates progressive stages of governance capability, ranging from informal AI deployment to optimized governance with continuous monitoring, explainability, and adaptive defense.

7. Discussion

7.1 Opportunities of Agentic AI in Cybersecurity

Agentic artificial intelligence has the potential to significantly transform modern cybersecurity operations because it combines reasoning, retrieval, tool use, and autonomous task execution within a single operational framework. Unlike traditional security automation systems that rely mainly on predefined rules or static workflows, agentic AI systems can interpret context, interact with external tools, retrieve relevant information, and adapt to changing operational environments. These capabilities create opportunities for improving efficiency, scalability, and responsiveness across cybersecurity operations.

One of the most important opportunities is accelerated threat detection. Security operations centers often process large volumes of alerts, logs, and threat intelligence feeds that exceed the capacity of human analysts. Agentic AI can assist by correlating alerts, retrieving contextual information, prioritizing suspicious activity, and recommending investigative actions in near real time. Retrieval-augmented generation and reasoning-capable models support this process by allowing AI systems to combine internal reasoning with external knowledge retrieval (Lewis et al., 2020; Yao et al., 2022).

Agentic AI can also reduce analyst workload. Many cybersecurity tasks are repetitive, time-consuming, and operationally intensive, including log analysis, ticket classification, incident documentation, vulnerability triage, and policy verification. Autonomous agents can automate portions of these workflows, allowing cybersecurity professionals to focus on higher-level analysis and strategic decision-making. This may improve operational efficiency while helping organizations address workforce shortages in cybersecurity operations. Incident triage is another important application area. Agentic AI systems can analyze alerts, gather supporting evidence, retrieve historical incident information, and prioritize incidents according to severity or operational

impact. This may improve response speed and reduce the time required for analysts to investigate routine events. Tool-connected language models and API-enabled agents make this possible because they can interact directly with security information and event management systems, cloud platforms, threat intelligence repositories, and ticketing tools (Patil et al., 2024; Schick et al., 2023).

Threat intelligence analysis also benefits from agentic AI capabilities. Cybersecurity teams frequently process large amounts of structured and unstructured threat intelligence data originating from reports, advisories, blogs, malware analyses, vulnerability disclosures, and security feeds. Agentic AI can summarize intelligence, identify relationships between indicators, extract key findings, and correlate emerging threats across multiple sources. Continuous retrieval and reasoning capabilities may therefore improve situational awareness and support proactive defense.

Continuous monitoring is another important opportunity. Since agentic AI systems can operate persistently across connected environments, they may support ongoing surveillance of security events, cloud infrastructure, user behavior, and system configurations. Continuous monitoring can improve detection of anomalies, unauthorized activity, and policy violations, especially in large enterprise environments where manual oversight alone may be insufficient.

Despite these benefits, the value of agentic AI depends heavily on governance, validation, and oversight. The same capabilities that enable efficiency and automation may also increase operational risk if agents operate without appropriate controls. Therefore, organizations should approach agentic AI as a tool for augmenting cybersecurity operations rather than replacing human expertise entirely.

7.2 Risks of Over-Automation

Although agentic AI offers operational advantages, excessive reliance on autonomous agents may introduce new cybersecurity vulnerabilities and governance risks. Over-automation occurs when organizations allow AI systems to operate with insufficient boundaries, validation, or oversight. In cybersecurity environments, this can create operational instability, inaccurate decision-making, and expanded attack surfaces.

One major concern is the reliability of AI-generated outputs. Studies on AI-assisted software development demonstrate that language models may generate insecure code, unreliable vulnerability fixes, or flawed recommendations (Pearce et al., 2023; Pearce et al., 2025; Perry et al., 2023). This means that autonomous agents operating in cybersecurity contexts may recommend incorrect remediation actions, misclassify threats, or introduce new weaknesses if outputs are not validated carefully.

Over-automation may also increase exposure to prompt injection and instruction manipulation. Greshake et al. (2023) demonstrate how indirect prompt injection can compromise LLM-integrated applications by embedding malicious instructions within retrieved or external content. If organizations allow agentic AI systems to process untrusted data without proper controls, attackers may manipulate agent behavior indirectly. Another important risk involves interaction with external packages, APIs, plugins, and third-party tools. Spracklen et al. (2025) show that code-generating language models may hallucinate software packages, creating opportunities for dependency confusion and malicious package exploitation. If autonomous agents are permitted to retrieve or install packages automatically, supply-chain risks may increase substantially.

Over-automation also raises concerns about operational transparency and accountability. If agents are allowed to make high-impact decisions independently, organizations may struggle to understand why actions occurred or who is responsible for failures. This becomes especially problematic when agents interact across multiple systems or environments without sufficient audit trails.

Human expertise therefore remains essential. Cybersecurity environments involve uncertainty, contextual interpretation, legal considerations, organizational priorities, and risk tradeoffs that AI systems may not fully understand. Human analysts are still necessary for validating critical decisions, reviewing unusual outputs, handling ambiguous situations, and ensuring that defensive actions align with organizational objectives. Rather than pursuing unrestricted autonomy, organizations should implement bounded autonomy in which agentic AI systems operate within defined permissions, monitored workflows, and escalation thresholds. This approach allows organizations to benefit from automation while reducing the risks associated with uncontrolled AI-driven operations.

7.3 Balancing Innovation and Control

The growing capabilities of agentic AI should not lead organizations to reject the technology entirely. Instead, organizations should focus on balancing innovation with governance and operational control. Agentic AI offers substantial opportunities for improving cybersecurity efficiency, responsiveness, and scalability, but safe deployment requires secure architecture, runtime governance, human oversight, transparency, and accountability.

A balanced approach begins with AI risk management. NIST's AI Risk Management Framework emphasizes identifying, measuring, managing, and governing AI-related risks throughout the system lifecycle (NIST, 2023, 2024). These principles are directly relevant to agentic AI because autonomous systems require ongoing evaluation rather than one-time approval.

OWASP's LLM security guidance provides additional protection by identifying major risks associated with prompt injection, insecure outputs, data leakage, and supply-chain vulnerabilities (OWASP Gen AI Security Project, 2025). Since agentic AI systems frequently depend on retrieval systems, external APIs, and connected workflows, these controls are important for preventing misuse and operational compromise.

Zero trust architecture also supports balanced deployment. Rose et al. (2020) explain that zero trust emphasizes continuous verification, least privilege access, and segmented environments. Applying these principles to agentic AI helps reduce the impact of compromised agents, unauthorized actions, or privilege escalation.

Balancing innovation and control also requires layered governance. Organizations should implement access control, tool validation, monitoring, auditing, escalation mechanisms, and human approval for sensitive actions. Runtime monitoring is especially important because agentic AI systems may behave differently across changing operational contexts.

Transparency and explainability are equally important. Analysts and decision-makers must understand how AI agents arrive at conclusions, which tools they use, and what information influences their outputs. Without explainability, organizations may either overtrust AI systems or reject them entirely.

The most effective approach is therefore controlled autonomy. Agentic AI should support cybersecurity teams through augmentation and operational assistance while humans retain authority over high-risk actions and strategic decision-making.

7.4 Implications for Organizations

The emergence of agentic AI has important implications for organizational cybersecurity governance, operational policy, and risk management. Organizations can no longer treat AI systems solely as analytical

tools or software utilities. Instead, AI agents should be treated as operational actors within enterprise environments.

This means that cybersecurity policies should explicitly address AI agent identities, permissions, audit trails, monitoring requirements, and accountability structures. AI agents should have defined operational scopes, role-based access permissions, and documented responsibilities comparable to those assigned to privileged human users or enterprise applications.

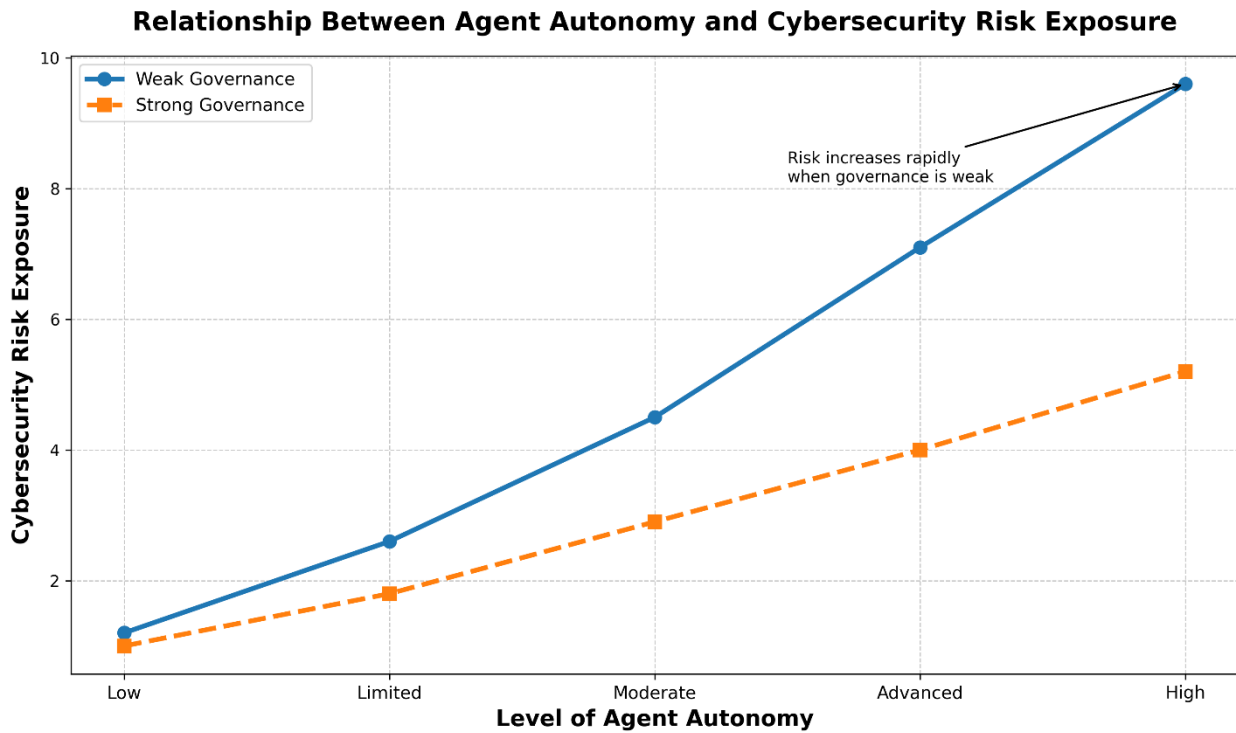
Organizations should also establish governance structures specifically for agentic AI deployment. These structures should define who is responsible for agent configuration, tool approval, monitoring, incident review, and operational oversight. Governance policies should include escalation procedures, approval requirements, and restrictions for high-risk activities.

AI auditing and model documentation are also increasingly important. Mitchell et al. (2019) emphasize the value of model cards for documenting system capabilities and limitations, while Raji et al. (2020) discuss internal algorithmic auditing for accountability. These practices can improve transparency, support compliance, and strengthen operational trust.

Lifecycle monitoring and post-market monitoring are equally necessary because AI systems may change over time due to evolving data, APIs, prompts, or operational environments. Mökander et al. (2022) demonstrate the importance of conformity assessments and post-deployment governance for maintaining accountability in AI systems. Organizations should therefore implement continuous review procedures rather than assuming that initial approval guarantees long-term safety.

Another important implication involves workforce readiness. Cybersecurity teams will increasingly need expertise in AI governance, prompt security, adversarial machine learning, runtime monitoring, and AI system auditing. Organizations may therefore need to update training programs, operational procedures, and governance frameworks to support AI-enabled cybersecurity environments.

Ultimately, responsible deployment of agentic AI requires both technical and organizational adaptation. The organizations most likely to benefit from agentic AI are those that integrate innovation with strong governance, operational discipline, and continuous oversight.



Conceptual analysis showing that cybersecurity risk exposure increases as AI agents gain higher autonomy and broader operational authority. Strong governance, runtime monitoring, and oversight mechanisms reduce overall risk exposure.

Relationship Between Agent Autonomy and Cybersecurity Risk Exposure

8. Recommendations

8.1 Technical Recommendations

Organizations deploying agentic AI in cybersecurity environments should adopt layered technical safeguards that reduce operational risk while maintaining system functionality. The first recommendation is to assign unique identities to all AI agents. Agents should operate through authenticated and traceable identities rather than shared credentials or unrestricted administrator accounts. This supports accountability and access governance.

Least privilege access should also be enforced consistently. Agents should only access the systems, APIs, data, and tools necessary for their approved tasks. Restricting permissions reduces the impact of compromised or manipulated agents and aligns with zero trust architecture principles (Rose et al., 2020).

Sandboxed execution environments are another important safeguard. High-risk agent operations should execute within isolated environments where unsafe actions can be contained without affecting production infrastructure. Segmented environments may also reduce lateral movement and operational exposure.

Continuous monitoring of agent behavior is essential. Organizations should monitor prompts, tool usage, outputs, accessed data, policy violations, and behavioral drift throughout the operational lifecycle. Runtime monitoring systems should detect anomalies, suspicious activity, and unauthorized actions before they escalate into major incidents.

Outputs generated by AI agents should also be validated before execution. Human review, confidence scoring, policy validation, and rule-based verification can reduce the likelihood of unsafe or inaccurate actions. Sensitive systems should require explicit approval before AI-generated recommendations are implemented.

Prompt injection detection and filtering mechanisms should be integrated into all tool-connected and retrieval-enabled systems. Since prompt injection remains one of the most serious risks affecting LLM-integrated applications, organizations should sanitize external inputs, isolate untrusted content, and enforce instruction hierarchy controls.

Comprehensive logging should be maintained across the entire AI workflow. Logs should capture prompts, tool calls, retrieved information, generated outputs, approvals, and execution results. These records support accountability, auditing, and forensic investigation.

Organizations should also validate software packages, APIs, plugins, and third-party tools before allowing agent interaction. This reduces exposure to dependency confusion, malicious packages, and supply-chain compromise.

Finally, AI systems should undergo adversarial security testing before deployment. Research on poisoning attacks, adversarial ML, extraction attacks, and privacy risks demonstrates that AI systems can be manipulated or compromised under hostile conditions (Biggio et al., 2012; Carlini & Wagner, 2017; Shokri et al., 2017; Tramèr et al., 2016). Regular red-team testing and adversarial evaluation therefore remain critical.

These recommendations collectively align with AI risk management principles, zero trust architecture, adversarial ML research, and modern LLM security guidance (NIST, 2023, 2024; OWASP Gen AI Security Project, 2025; Rose et al., 2020).

8.2 Governance Recommendations

Technical controls alone are not sufficient for governing agentic AI systems. Organizations should establish formal governance structures that define accountability, oversight, operational boundaries, and lifecycle management procedures.

The first recommendation is to establish an agentic AI governance policy that clearly defines acceptable use, restricted actions, approval procedures, monitoring requirements, and escalation mechanisms. Policies should also define operational roles and responsibilities related to AI deployment, monitoring, and incident response. Organizations should define accountability for agent actions explicitly. Governance frameworks should identify who is responsible for agent configuration, tool integration, monitoring, oversight, and approval of high-risk operations. This helps reduce accountability gaps associated with autonomous systems.

Human approval should remain mandatory for high-risk tasks involving sensitive infrastructure, security policy changes, financial systems, production environments, or privileged operations. Human oversight helps reduce operational risk and improves accountability.

Regular AI security audits should also be conducted. Audits should evaluate compliance with organizational policies, AI governance standards, runtime monitoring requirements, and operational security controls. Internal algorithmic auditing may help identify weaknesses before major incidents occur (Raji et al., 2020).

Governance structures should align with NIST AI RMF, zero trust architecture, and relevant cybersecurity frameworks. Aligning deployment practices with recognized standards improves governance consistency and supports regulatory compliance.

Organizations should also develop incident response plans specifically for AI-related failures or misuse. Incident response procedures should define containment actions, escalation workflows, rollback procedures, forensic investigation requirements, and post-incident review processes.

Model documentation and activity records should be maintained continuously. Documentation may include model capabilities, operational limits, approved tools, known risks, monitoring requirements, and governance

decisions. Activity records should preserve evidence of prompts, decisions, approvals, and system interactions.

Post-deployment monitoring and periodic conformity assessments are equally important because AI systems may evolve over time. Mökander et al. (2022) emphasize that ongoing monitoring is necessary for maintaining accountability and regulatory alignment in AI systems.

Together, these governance recommendations support secure, transparent, and accountable deployment of agentic AI systems.

8.3 Research Recommendations

Although agentic AI offers important opportunities for cybersecurity, several research gaps remain unresolved. Future research should therefore focus on improving the security, explainability, governance, and operational reliability of autonomous AI systems.

One important area is agentic AI red teaming. Researchers should develop structured methods for testing autonomous agents against prompt injection, adversarial manipulation, unsafe tool use, privilege escalation, and operational compromise. Existing adversarial ML research provides a useful foundation, but agentic AI introduces additional complexities because agents can interact with tools and external systems.

Explainable agent behavior is another important research area. Current language models often provide limited transparency regarding how decisions are produced or why specific actions are recommended. Improved explainability methods may strengthen analyst trust, auditability, and regulatory compliance.

Autonomous cyber defense benchmarking also requires further development. Researchers should establish standardized benchmarks for evaluating detection quality, response speed, operational safety, and governance effectiveness in agentic cybersecurity systems.

AI identity governance and runtime containment models are equally important. Future studies should examine how organizations can manage AI identities, limit permissions dynamically, and contain compromised or malfunctioning agents during runtime operations.

Prompt injection defense remains a critical challenge. Since prompt injection can manipulate reasoning-capable systems through both direct and indirect methods, future research should improve prompt filtering, instruction hierarchy enforcement, and context validation mechanisms.

Secure tool-use evaluation is another important area because tool-connected agents may interact with APIs, cloud services, software packages, and enterprise systems. Research should therefore examine safe API governance, permission management, runtime validation, and tool trustworthiness.

Package hallucination mitigation also requires additional study. Since code-generating models may recommend nonexistent or malicious packages, future work should examine methods for validating dependencies and preventing supply-chain exploitation (Spracklen et al., 2025).

Research should continue examining data poisoning, model extraction, memorization, membership inference, and adversarial robustness in AI-enabled cybersecurity systems (Carlini et al., 2021; Carlini et al., 2022; Geiping et al., 2020; Jagielski et al., 2018; Papernot et al., 2018; Shokri et al., 2017; Tramèr et al., 2016). These risks remain highly relevant because agentic AI systems often rely on large-scale data access, retrieval systems, and connected operational environments.

9. Conclusion

Agentic artificial intelligence is becoming an important development in modern cybersecurity because it combines reasoning, retrieval, tool use, and autonomous task execution within a single operational framework. These capabilities create opportunities for improving cyber defense by accelerating threat detection, supporting threat intelligence analysis, improving incident triage, enabling continuous monitoring, and reducing operational workload for cybersecurity teams.

However, the same autonomy that creates operational value also introduces significant cybersecurity and governance risks. Agentic AI systems may be vulnerable to prompt injection, insecure tool use, privilege escalation, data leakage, supply-chain compromise, unsafe automation, insecure code generation, and accountability failures. These risks become more serious when agents operate across connected systems, external APIs, cloud environments, or sensitive enterprise infrastructure.

The study therefore argues that safe deployment of agentic AI requires a layered defensive framework rather than isolated security controls. Effective governance should combine agent identity management, secure tool governance, prompt and input security, runtime monitoring, behavioral analytics, human oversight, auditability, explainability, and incident response. Continuous monitoring and lifecycle governance are especially important because agentic AI systems may evolve after deployment due to changing data, APIs, retrieval systems, or operational environments.

The findings also emphasize that agentic AI should augment cybersecurity operations rather than replace human expertise entirely. Human oversight remains necessary for validating high-risk actions, interpreting ambiguous situations, and maintaining accountability in security-critical environments.

Overall, responsible deployment of agentic AI depends on balancing innovation with governance and operational control. AI risk management frameworks, generative AI security guidance, zero trust architecture, algorithmic auditing, and post-deployment monitoring provide important foundations for achieving this balance (Mitchell et al., 2019; NIST, 2023, 2024; OWASP Gen AI Security Project, 2025; Raji et al., 2020; Rose et al., 2020).

References

1. NIST. (2023). AI Risk Management Framework. Artificial Intelligence Risk Management Framework (AI RMF 1.0), 1(1). <https://doi.org/10.6028/nist.ai.100-1>
2. NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. <https://doi.org/10.6028/nist.ai.600-1>
3. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture. NIST special publication, 800(207), 1-52.
4. Marasani, Y. (2025). Explainable AI Frameworks for Patient-Level Claims Data Analytics. J Artif Intell Mach Learn & Data Sci 2025, 8(1), 3382-3390.
5. OWASPGenAIProject Editor. (2025, April 28). OWASP Top 10 for LLM Applications 2025. OWASP Gen AI Security Project. <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
6. ENISA Threat Landscape 2025 | ENISA. (2025, November 6). Europa.Eu. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025>
7. Securing Machine Learning Algorithms | ENISA. (2024, February 21). Europa.Eu. <https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms>

8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
9. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)* (pp. 2633-2650).
10. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2025). Asleep at the keyboard? assessing the security of github copilot's code contributions. *Communications of the ACM*, 68(2), 96-105.
11. MARASANI, Y. (2023). Machine Learning Models for Predicting Patient Treatment Switching Using Claims Data. *Frontiers in Computer Science and Artificial Intelligence*, 2(1), 59-66.
12. Pearce, H., Tan, B., Ahmad, B., Karri, R., & Dolan-Gavitt, B. (2023, May). Examining zero-shot vulnerability repair with large language models. In *2023 IEEE symposium on security and privacy (SP)* (pp. 2339-2356). IEEE.
13. Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37, 126544-126565.
14. ALAMPALLY, J. (2024). Enhancing data quality and trust in AI systems through robust data engineering. *Frontiers in Computer Science and Artificial Intelligence*, 3(1), 120-130.
15. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
16. Vallemoni, R. K. (2021). Settlement, Fees, and Interchange: Data Models for Accurate Reconciliation and Exception Handling. AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT.
17. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36, 68539-68551.
18. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security* (pp. 79-90).
19. Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023, November). Do users write more insecure code with ai assistants?. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security* (pp. 2785-2799).
20. Peng, Z., Wang, Z., & Deng, D. (2023). Near-duplicate sequence search at scale for large language model memorization evaluation. *Proceedings of the ACM on Management of Data*, 1(2), 1-18.
21. Li, Y., Liu, Y., Deng, G., Zhang, Y., Song, W., Shi, L., ... & Wang, H. (2024). Glitch tokens in large language models: Categorization taxonomy and effective detection. *Proceedings of the ACM on Software Engineering*, 1(FSE), 2075-2097.
22. ALAMPALLY, J. (2024). Real-Time and Near-Real-Time Analytics in Healthcare Data Ecosystems. *Journal of Computer Science and Technology Studies*, 6(1), 314-324.
23. Spracklen, J., Wijewickrama, R., Sakib, A. N., Maiti, A., & Viswanath, B. (2025). We have a package for you! a comprehensive analysis of package hallucinations by code generating {LLMs}. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 3687-3706).

24. Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., ... & Duan, N. (2024). Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing*, 3, 0063.
25. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
26. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
27. Vallemoni, R. K. (2022). Canonical payment data models for merchant acquiring: Merchants, terminals, transactions, fees, and chargebacks. *International Journal of Computer Science and Engineering (ISCSITR-IJCSE)*, 3(1), 42-66.
28. KOTA, S. K. (2022). A Real-World Deployment of an Enterprise Conversational AI Platform for Demand Generation and Lead Generation Using Guided Workflows with a Rasa-Based Chatbot. *Frontiers in Computer Science and Artificial Intelligence*, 1(1), 24-30.
29. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.
30. Nagraj, A. (2024). GraphQL in Wealth Management Platforms: Optimizing Data Access and Performance. *British Journal of Multidisciplinary Studies*, 2(1), 16-24.
31. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (mlops): Overview, definition, and architecture. *IEEE access*, 11, 31866-31879.
32. Chou, E., Tramer, F., & Pellegrino, G. (2020, May). Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)* (pp. 48-54). IEEE.
33. Papernot, N., Mcdaniel, P., Sinha, A., & Wellman, M.P. (2018). SoK: Security and Privacy in Machine Learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399-414.
34. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)* (pp. 601-618).
35. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506-519).
36. Biggio, B., Nelson, B., & Laskov, P. (2012, June). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1467-1474).
37. Vallemoni, R. K. (2022). Authorization-to-settlement at scale: A reference data architecture for ISO 8583/ISO 20022 coexistence. *Journal of Computer Science and Technology Studies*, 4(1), 88-98.
38. Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I., Saini, U., ... & Xia, K. (2008). Exploiting machine learning to subvert your spam filter. *Leet*, 8(1-9), 16-17.

39. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 19-35). IEEE, Institute of Electrical and Electronics Engineers.
40. Geiping, J., Fowl, L., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., & Goldstein, T. (2020). Witches' brew: Industrial scale data poisoning via gradient matching. arXiv preprint arXiv:2009.02276.
41. Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 IEEE symposium on security and privacy (sp) (pp. 39-57). Ieee.
42. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.
43. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022, February). Quantifying memorization across neural language models. In The Eleventh International Conference on Learning Representations.
44. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021, May). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
45. Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, 32(2), 241-268.