INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
ISSN: 3106-7980: 3106-7972
Volume 3, Issue 3, 2026

SECURING AGENTIC ARTIFICIAL INTELLIGENCE SYSTEMS: ASSESSING PROMPT INJECTION, MEMORY POISONING, AND AUTONOMOUS TOOL-USE RISKS IN INTELLIGENT APPLICATIONS

Faheem Ahmed[1], Dr Hussain Shah[2], Nazia Azim[3], Faisal Rahman[*4]

[1]*Assistant Professor, Department of Information Technology, University of Sindh, Jamshoro, Pakistan*
[2]*Lecturer, Shaikh Zayed Islamic Centre University of Peshawar, Pakistan*
[3]*Lecturer, Department of Computer Science Abdul Wali khan university Mardan*
[*4]*Information Technology Management, Southwest Baptist University MO USA*

[1]faheem.abbasi@usindh.edu.pk, [2]safihussain@uop.edu.pk, [3]n.azim@awkum.edu.pk,
[*4]faisal.rahman@sbuniv.edu

DOI: https://doi.org/10.5281/zenodo.21718185

Keywords
Agentic Artificial Intelligence, Prompt Injection, LLM Security Memory Poisoning, Autonomous AI Agents

Article History
Received: 28 May 2026
Accepted: 01 July 2026
Published: 31 July 2026

Corresponding Author: * Faisal Rahman

Abstract

Artificial Intelligence (AI) has come a long way, and the advent of Agentic Artificial Intelligence systems goes beyond traditional AI offerings, such as reasoning, planning, memory retention, and access to external tools. These are areas that offer important potential for automation and intelligent decision making, but they also bring along with them complex cyber security challenges. Unlike conventional software systems, Agentic AI applications are based on flexible interactions between LLMs, memory modules, external data sources, and autonomous mechanisms for tool usage, which introduces new attack vectors. Some of the most significant emerging risks include unauthorized autonomous tool execution, memory poisoning, and prompt injection attacks, all of which can erode trust in AI systems, data security, and organizational reliability. The goal of this study was to understand the insights of cybersecurity practitioners on Agentic Artificial Intelligence-related security threats. A qualitative, phenomenological study approach was used to explore experts' lived experiences and interpretations of AI security challenges. The data were obtained using semi-structured interviews conducted with 18 professionals such as AI security researchers, cybersecurity experts, machine learning engineers, and software architects. Thematic analysis method by Braun and Clarke was used to analyse the collected data to find any significant pattern and theme. The results pointed to four key themes: (1) Agentic AI is a growing cybersecurity attack vector, (2) prompt injection is a key threat to AI reliability, (3) memory poisoning is a challenge to long-term trust in AI, and (4) autonomous tool usage is a governance and control issue. The participants highlighted the limitations of conventional cybersecurity methods to defend autonomous AI systems, given their adaptive nature, intricate structures, and self-reliant execution. The research emphasises the need for AI-specific security models, least-privilege access control, memory validation, continuous monitoring, and oversight. The study adds to the growing area of AI cybersecurity by offering empirical findings on the experiences of experts and actionable recommendations for securing next-

generation autonomous AI applications. The results are relevant to policymakers, organizations, cybersecurity experts, and developers of Agentic AI technologies, as they aim to encourage secure and responsible use of these new technologies. Agents are exploring the possibility of monitoring the user's input and the resulting output.The agents are investigating potential monitoring of the input and output of the AI system.

INTRODUCTION

1.1 Background of the Study

Artificial Intelligence (AI) is one of the most impactful technological advancements of the 21st century, revolutionizing how organizations handle information, automate processes and make decisions. Over the years, AI has progressed to become more autonomous and adaptable, ranging from early rule-based expert systems to today's sophisticated deep learning architectures. The combination of recent developments in large language models (LLMs), generative AI, and autonomous reasoning systems has greatly broadened the capabilities of intelligent applications. Unlike legacy software systems, which run on a set of instructions, today's AI systems are more and more able to read complex goals, write answers, create plans, and communicate with the outside world.

Agentic AI is a new paradigm in AI evolution. Agentic AI is the use of AI tools that can be self-sufficient in their analytical, strategic, action, memory, and interaction functions with external tools or digital environments. They are made up of several modules like LLM, reasoning, memory, planning, and external tool integration, which can handle complex tasks with minimal human intervention.

Agentic AI systems are more proactive than traditional generative AI systems, which are mainly limited to reacting to user requests. In the case of a software development AI agent, it can analyze project requirements, write code, test solutions, detect bugs, and interact with external development tools. In the finance sector, AI agents can also analyze market data, create reports, and provide guidance on decision-making. With regard to healthcare, agentic AI can help healthcare professionals by analyzing patient data and suggesting potential interventions. These systems can handle multiple-step tasks autonomously, providing notable opportunities for enhancing efficiency, productivity, and decision-making in various industries.

But with the rise of autonomy of AI-based systems, unprecedented cybersecurity challenges have emerged. Traditional solutions for cybersecurity were primarily developed to safeguard traditional software applications, databases and network infrastructure where the mode of operation was predictable and rule-based. One of the major distinctions between agentic and non-agentic AI systems lies in their complexity and ability to apply sophisticated reasoning, probabilistic results, memory, and external information. This makes securing these systems a new challenge in cybersecurity, as it involves comprehending the risks that arise from AI autonomy itself.

Prompt injection attacks are among the most important security risks related to agentic AI. Prompt injection is a technique used by attackers to trick the AI by altering the prompts to cause it to act in ways that are not intended or secure. Prompt injection attacks are different from standard cyberattacks, which focus directly on vulnerabilities in the system, as they make the most of the interpretative nature of AI models. An attacker can embed hidden commands in documents, webpages, databases or user inputs that give instructions to an AI agent that it never intended to carry out. This is especially risky if the AI agents have access to sensitive data or external tools.

For instance, a malicious instruction could be placed inside a file that an AI agent handling organisational documents is unaware of, and then it discovers the information that should not be shared. Likewise, an external application can be tricked into making unauthorized actions via an AI assistant. These scenarios illustrate that

prompt injection is not just a technical issue, but also one of understanding the reasoning behind prompts and prioritizing instructions in AI.

Memory poisoning is another new threat that is becoming a matter of concern in the security field. Advanced AI agents often incorporate memory capabilities to recall past interactions, remember user preferences, and make more informed decisions in subsequent interactions. Memory increases the attack surface while also increasing personalisation and efficiency. Memory poisoning happens when attackers intentionally feed false or misleading or malicious information into an AI agent's memory. This tainted data can affect future actions and choices and lead to continued, negative behaviour.

Memory poisoning is especially critical in retrieval-augmented generation (RAG) systems and in enterprise AI deployments where agents are constantly ingesting and pulling information from their internal knowledge sources. Memory poisoning differs from temporary prompt manipulation, in that the malformed information can subsequently affect future interactions, thereby having potentially long-lasting effects. As a result, safeguarding the integrity and validation of memory during the development of secure agentic AI has become paramount.

Another challenge is the use of autonomous tools. Today, AI agents are more and more specifically created to communicate with exterior tools, akin to APIs, databases, software program platforms, and automated services. Integration with tools greatly enhances the capabilities of AI, but it also expands attack surfaces. Without proper permissions, AI agents could gain access to resources that they shouldn't, execute unauthorized commands, or accidentally do malicious things. The challenge is greater when AI systems are given general freedom of action with inadequate supervision or human control.

Researchers, tech companies and cybersecurity institutions are all interested in the security issues surrounding agentic AI. Specialized security approaches are emphasized by frameworks like the National Institute of Standards and Technology (NIST) AI Risk Management Framework, MITRE ATLAS, and the OWASP Top 10 for Large Language Model Applications, highlighting the need for targeted strategies to tackle AI-specific vulnerabilities. The principles such as secure AI development and continuous monitoring, risk assessment, and human-centred governance are emphasized within these frameworks.

While the awareness of AI security has been growing, current research on AI security is predominantly limited to technical evaluation, demonstration of vulnerabilities and attacks on the model level. Empirical studies of cybersecurity professionals' awareness and experience concerning the practical issues of securing agentic AI systems are scarce. There is added value in the expert perspective because the cybersecurity practitioner is directly experiencing challenges with implementation, organizational challenges, and new attack patterns.

This is particularly significant in the context of emerging technological landscapes, such as Pakistan, where organizations and institutions are increasingly adopting AI-driven systems. With AI agents being introduced into various aspects of business, education, healthcare, and government services, it is crucial to understand the security consequences of their integration into ensure their responsible use. However, there is a dearth of research that discusses AI security perceptions in the field of such contexts.

Hence, a qualitative phenomenological approach is used in this study to delve into the views of cybersecurity experts on the security issues of agentic AI systems. This study will investigate expert experiences with prompt injection, memory poisoning and autonomous tool usage risks to gain insights into vulnerabilities and protection measures needed for the security of intelligent autonomous applications.

1.2 Problem Statement

While the burgeoning growth of agentic AI systems has ushered in fresh possibilities for automation and intelligent decision-making, it has also brought with it intricate cybersecurity challenges. Current security solutions used for conventional software applications may not be sufficient to tackle the vulnerabilities arising with

the autonomous reasoning, memory-based learning, and tool interaction capabilities.

While some studies have analysed vulnerabilities of AI from a technical angle, there has been little work that has explored how cybersecurity professionals view and manage these vulnerabilities in practice. However, there is a gap in understanding on practical challenges in obtaining agentic AI applications, as well as a lack of qualitative evidence of experts' experiences.

To fill this gap, this study examines cybersecurity experts' perceptions of prompt injection, memory poisoning, and autonomous tool-use vulnerabilities.

1.3 Research Objectives

The study aims to:

Consider the experiences of security professionals about the security issues faced in agentic artificial intelligence systems.Discuss with security professionals their experiences with security problems in agentic AI systems.

A. Discuss what experts think about prompt injection vulnerabilities.

3. Explore how memory poisoning affects trustworthiness of an AI agent.

4. Discuss security issues related to using tools independently.

5. Determine ways to enhance AI agent security.

1.4 Research Questions

RQ1: What are the security issues that cybersecurity professionals face when using AI agents?

RQ2: What are the risks and vulnerabilities of prompt injection attacks?

RQ3: What is memory poisoning and how can it impact the reliability and trustworthiness of AI agents?

RQ4: How can the security practices minimize the risks associated with autonomous tool use in Agentic AI?

The purpose of this study is to investigate the significance of the study.

This study adds to the current growing body of work in the field of AI Cybersecurity by offering qualitative information on the perspectives of experts on the protection of autonomous intelligent systems. The results could help AI creators, cybersecurity experts, organisations and policy makers in creating secure AI architectures.

Moreover, this study emphasises the need to transcend conventional cybersecurity strategies and create security frameworks tailored to the unique nature of AI tools, particularly concerning autonomy, memory integrity, and responsible use.

2. LITERATURE REVIEW

2.2 Agents & Artificial Intelligence Systems

The evolution of Artificial Intelligence technology has evolved from rule-based systems to highly autonomous intelligent applications that are capable of reasoning, learning, and interacting with complex environments. The advent of Large Language Models (LLMs), including those based on GPT, and other foundation models has paved the way for this change by giving machines the ability to grasp and understand natural language, respond in a human-like manner, and undertake advanced cognitive tasks. Recently, however, the path of AI research is shifting from generative models to Agentic Artificial Intelligence (AI) systems that are engineered to act independently, planning their actions, retaining information, and interacting with external tools (Xi et al., 2023).

Agentic AI systems are a fusion of various technologies, such as large language models (LLMs), reasoning systems, planning components, memory designs, and tool usage capabilities. AI agents are capable of autonomously figuring out the order of actions needed to accomplish particular goals, whereas conventional AI systems just make predictions or create responses based on the consumer's instructions. This stand-alone functionality has opened up a lot of avenues for automation in fields like cybersecurity operations, healthcare management, financial services, education, and enterprise decision-making (Wang et al., 2024).

Agentic AI brings a new paradigm to the security landscape. A traditional software application is based on deterministic rules, that is, the behaviour of the application can be predicted and controlled by preprogrammed logic. Agentic

AI systems, on the other hand, take action based on a blend of probabilistic reasoning, context, and built-up knowledge. Such dynamic behavior gives rise to other attack surfaces that traditional cybersecurity approaches can't detect (NIST, 2023).

The researchers have pointed out that AI agents pose dangers at various stages, such as model behavior, instruction processing, memory management and external interaction. With the ability to independently perform tasks, security breaches could not only compromise information confidentiality but also result in inappropriate decision-making or actions by the AI agents. If manipulated with malicious instructions, for instance, an AI agent joined with enterprise systems can accidentally reveal private enterprise data.

As a result, there is a growing need for dedicated security structures to handle the agentic AI's.This has thus led to the demand for dedicated security structures to deal with the agentic AIs. Security principles for AI systems should take into account vulnerabilities specific to AI, not just traditional cybersecurity practices, according to organizations like the Open Worldwide Application Security Project (OWASP) and the National Institute of Standards and Technology (NIST). These methods involve risk management, secure model building, access control, monitoring, and human checking systems.

While there has been growing concern about security issues related to AI, the majority of the research so far focuses on technical analysis of vulnerabilities. There is relatively little research that examines the practical challenges that cybersecurity practitioners face in securing autonomous AI systems, their perceptions of those challenges, and how they react to them. This underscores the importance of studying AI security issues qualitatively to understand the expert viewpoints on these new risks.

2.3 Ethical Implications of AI Systems.2.4 AI Systems and the Ethics of Misuse.

One of the most critical security threats of LLM and agentic AI is prompt injection. Prompt injection attacks target the ability of AI models to follow instructions, in contrast to traditional attacks that exploit software vulnerabilities or network vulnerabilities. The attacks are designed to trick the input instructions given to AI systems into yielding unwanted results or taking unauthorized action (Greshake et al., 2023).

There are two types of prompt injection attacks: direct prompt injection and indirect prompt injection. Direct prompt injection is when an attacker sends out malicious commands through the user. For instance, an attacker could tell an AI system to bypass previous security rules or even expose private information. Indirect Prompt Injection: malicious techniques are included in external content, like websites, documents, emails or databases, that are accessed by the AI agent during the execution of the task.

Prompt injection poses a significant threat in agentic AI systems, where AI agents may have various other functionalities besides text generation. Combined with external tools, APIs or databases, manipulated instructions can lead agents to take actions that they were not intended to take. For example, if an AI assistant has access to company records, it might be possible for it to be used to leak confidential data or to alter critical information.

Prompt injection has been shown to be a fundamental challenge as LLMs are not always reliable to differentiate between trusted and untrusted prompts. In contrast to typical application scenarios, which clearly distinguish between commands and data, AI models process instructions and information via natural language interaction, which makes it hard to set clear security boundaries (Perez & Ribeiro, 2022).

There are several studies that have stated the need to create defense against prompt injection. There are four approaches suggested: instruction hierarchy management, input filtering, output monitoring, sandboxing, and human-in-the-loop verification. But, researchers believe complete prevention is difficult due to the intrinsic nature of the design principles of language-based AI systems (Zhang et al., 2024).

Prompt injection is a significant issue for agentic AI applications, as successful exploitation can impact not just the agent's output, but also its

independent decision-making. Thus, it is crucial to learn from the experts' experiences on prompt injection threats to be able to formulate adequate security solutions.

2.3 Poisoning risks of autonomous AI agents.

One of the key characteristics distinguishing agentic AI systems from traditional generative AI models is their ability to remember.Unlike standard generative AI systems, one of the defining characteristics of agentic AI is their memory. AI agents can have short-term memory and long-term memory to remember past interactions, user preferences, and enhance future performance. Memory can be used to make the system more personalized and efficient, but also new security vulnerabilities are created.

Memory poisoning is the deliberate introduction of incorrect, false, or malicious information into an AI system's memory. In future interactions, the AI agent might provide faulty answers or make poor choices based on the poisoned data it retrieved. Memory Poisoning can generate long-term vulnerabilities as corrupted data might still impact the agent over time (Zhang et al., 2024), unlike temporary attacks which only affect one interaction.

The use of Retrieval-Augmented Generation (RAG) architectures is crucial in systems where memory poisoning can play a significant role. In the RAG-based systems, the AI models fetch data from outside knowledge bases before creating the answer. Attackers can gain access to these knowledge repositories and manipulate the information that AI agents have access to. This poses challenges for businesses that depend on AI systems for making decisions, providing customer service, or managing knowledge.

One of the problem areas with memory security is the difficulty in telling the difference between legitimate information and malicious information. However, while AI agents are supposed to learn from interactions, the challenge of ensuring that they don't learn harmful things but still remain useful remains a technical issue. To safeguard AI memory systems, researchers highlight the importance of implementing memory validation methods, source verification, access controls, and ongoing monitoring.

But from a cybersecurity point of view, memory poisoning poses a threat to traditional notions of data security. In traditional systems, information protection primarily covers encryption and access control. AI memory, however, needs extra safeguards to ensure that it is not only confidential, but also accurate, reliable, and contextually appropriate.

While vulnerabilities to memory have been studied in a technical context, there is not much research on the understanding of memory poisoning implications by cybersecurity professionals. The views of experts can be valuable in understanding the readiness of the organisation, implementation challenges and effective mitigation measures.

2.4 Autonomous Tool-Use Vulnerabilities

A key advantage of agentic AI systems is their ability to communicate with the outside world and digital environments. AI agents can browse databases, run software commands, fetch data from the internet and interact with other applications via APIs. This feature increases their utility, but also adds to the cyber security challenges.

When the tools are used autonomously, there are issues of over permission, unauthorized access and unwanted actions. If you give AI agents too much power, it could lead to significant problems if it's being used maliciously or if there's a mistake. An AI agent, therefore, linked into financial systems could be capable of conducting illegal transactions if adequate protections aren't in place.

The principle of least privilege (PoLP) is a fundamental security principle for AI agents that has been identified. Access to minimal resources should be granted to AI systems to perform tasks assigned. Also, tool interactions need to be tracked using logging and/or approval procedures to eliminate the risk of undesirable autonomous actions.

One other issue is the communication among agents. In the future, ecosystems of AI may have several autonomous agents working together.

This can help to streamline the process, but it also can create complicated security landscapes in which compromised agents can impact other systems. Trusted, authenticated, and secure communication between AI agents will be more crucial than ever.

The self-sufficiency of the use of tools also presents governance-related issues. Businesses need to decide on the right amount of human oversight and set clear limits on what decisions an AI system can make without human input.Companies should decide on suitable amounts of human supervision and limit decisions made by AI systems without human intervention. This highlights the importance of human-centered AI governance models.

2.5 AI Security Governance and Risk Management

As AI systems become more complex, the need for frameworks governing responsible AI security has been growing, with international bodies responding to the challenge. The NIST AI Risk Management Framework focuses on core AI risk management principles like reliability, transparency, accountability, and continuous risk management. Likewise, OWASP has found a handful of vulnerabilities with LLM, such as prompt injection, insecure plugins, and over-agency.

Achieving AI security governance will be a collaborative effort between researchers, developers, cybersecurity experts, and policymakers. Unfortunately, many potential risks include organizational decisions, ethical issues, and human behavior, which cannot be solved with technical solutions alone.

Governance is especially critical with agentic AI systems, as their autonomy is accompanied by greater capabilities and responsibilities. There should be clear lines of accountability for AI decisions, security monitoring, and incident response within organisations.Organisations should have clear lines of accountability for AI decisions, security monitoring, and incident response.

2.6 Research Gap

As described in the literature, there are some technical gaps existing in the literature about vulnerabilities of AI. First, most previous studies are experimental security tests, and not on understanding the real world experiences of cybersecurity professionals. Second, although there has been some qualitative work on this topic, it has been limited to how experts understand the risks related to prompt injection, memory poisoning, and autonomous usage. Thirdly, the research on agentic AI security is still mainly from advanced technological countries, and the viewpoints of the developing technological ecosystem are still limited.

Moreover, as more and more organizations start using autonomous AI tools, it is crucial to be aware of real-life security concerns. Technical frameworks will not necessarily address organizational restrictions, implementation challenges, and professional experiences of AI security.

To fill this missing link, the study examines the reality of the cybersecurity experts' experiences and perceptions of the security challenges created by agentic AI systems. The study is a qualitative inquiry which seeks to gain deeper knowledge of the vulnerabilities that might arise and to find practical ways of securing autonomous intelligent applications.

3. METHODOLOGY

3.1 Research Paradigm and Research Approach

This study chose an interpretivist research paradigm as the key focus and goal was to interpret cybersecurity professionals' perceptions, experiences, and interpretations of the new security challenges that Agentic Artificial Intelligence systems introduce. Interpretivism adopts the view that reality is socially created by individual experiences, and that complex phenomena can be better understood through examining participants' perspectives in their professional settings (Creswell & Poth, 2023).

Agentic AI security is a fledgling research field where “real-world” experience and expertise is growing, making qualitative study a suitable method. While quantitative methods can

highlight trends or patterns in variables, they might not adequately capture the experiences, interpretations, and reactions of professionals facing complex AI security issues. So this study was inclined towards gaining deep and detailed understanding of experts regarding the concepts of prompt injection, memory poisoning, and autonomous tool use risks.

3.2 Research Design

This study used a qualitative phenomenological research design. In the phenomenology approach, the emphasis is on investigating how people experience a phenomenon and the significance they give it. The phenomenon studied in this research was the experience of protecting Agentic AI systems from new cybersecurity threats.

The phenomenological approach was chosen because Agentic AI security is a dynamic field, and practitioners are continually facing novel vulnerabilities, implementation hurdles, and risk-management issues. In this study, the researchers sought to gain insights into the practical challenges of making autonomous AI applications a reality by immersing themselves in those experiences.

The study took an exploratory approach, due to the scarcity of empirical research on professional experience in the use of Agentic AI security. The method enabled the participants to share their observations, possible issues and recommendations from a professional perspective.

The third component of the research, Participants and Sampling Strategy, is detailed below.

This study's participants included cybersecurity and artificial intelligence practitioners with first-hand experience in AI systems, cybersecurity operations, software development, or AI-driven applications.

The participants were chosen using a purposive sampling technique, as the study needed people who have extensive knowledge and experience related to the topic of AI security. With purposive sampling, researchers can select those who have information-rich lives and can give them meaningful information about what they are studying.

The total number of participants in the study were 18. The participants included representatives from a wide-ranging group of professions, such as:

- AI security researchers
- Cybersecurity professionals
- Machine learning engineers
- Software developers
- AI system architects
- Information security consultants

This diversity of participants enabled the study to reflect several different viewpoints on the vulnerabilities of Agentic AI and security measures.

Table 1

Sample characteristics of participants in the study.

Participant Code: Professional Background: Experience Level:

E1. AI Security Researcher: 5+ years' experience.E1. AI Security Researcher: 5+ years experience.

E2 – Cybersecurity Analyst (6 years)

Programmer/Analyst / E3 Machine Learning Engineer (4 years)

E4 Software Architect - 8 years experience

E5 Cybersecurity Consultant – 7 years

E6 AI Developer 5 years

E7 Information Security Specialist 6 years

E8 AI Research Engineer (4 years)

E9 - Cybersecurity Researcher - 5 Years

E10 - Cloud Security Engineer 7 years

AI Systems Developer (E11) - 6 years

E12 Security Operations Specialist 5 years

E13	Machine Learning Researcher	4 years
E14	Penetration Testing Specialist	8 yrs.
E15	AI Governance Professional	6 years
E16	Software Security Engineer (5 years)	
E17	Cyber Risk Analyst 7 Years	
A21	AI Engineer (AI R&D) 2 years	

3.4 Data Collection Procedure

Semi-structured interviews were used to collect data as it gives the interviewer flexibility and consistency across interviewees. Semi-structured interviews can be used to probe a set of predetermined topics, as well as to elicit surprising experiences and insights from the interviewee.

The interviews took between 40 – 60 minutes and were held online via communication platforms as well as professional meetings. Participants received information regarding the purpose of the study, the confidentiality procedures and rights of voluntary participation before conducting the interviews.

The interview questions were related to four major areas:

First-hand knowledge of Agentic AI systems.Personal experience with Agentic AI systems.

Prompt injection attacks present a security threat.

4. How to prevent memory poisoning.5. How to avoid memory poisoning risks.
5. Autonomous AI systems and the problems they bring on user autonomy and security governance.6. Problems with using AI tools autonomously and on security governance.

Contestants were asked to share comprehensive answers, illustrations, and expert insights concerning the issues of AI security.

3.5 Interview Protocol

The semi-structured interviews followed the questions listed below:

Understand the concept of Agentic AI.General understanding of Agentic AI.

Q1. What do you think Agentic Artificial Intelligence systems are and how are they different from traditional AI applications?

Q2. What are some of the security concerns you've seen when using autonomous AI?

Prompt Injection Risks

Q3. What are your thoughts on prompt injection attacks as a security threat to AI agents?

Q4. Agentic AI systems can be susceptible to instruction tampering due to several factors.There are several factors that render Agentic AI systems susceptible to instruction tampering.

Q5. How can organizations minimize prompt injection threats?

Memory Poisoning Challenges

Q6. What challenges could arise in applying memory-based AI systems to create new cybersecurity vulnerabilities?

Q7. What are some of the potential hazards of introducing bad data into an AI agent's memory?

Autonomous Tool-Use Risks

Q8. What security issues can be created when AI agents can communicate with external tools and APIs?

Q9. How to keep organisations in control of autonomous AI actions?

Future Security Practices

Q10. What are your recommendations for securing future Agentic AI Systems?

3.6 Data Analysis Procedure

Braun and Clarke's (2006) six phases of thematic analysis were used to analyze the interview data collected. Thematic analysis was chosen as a suitable method of analysis because it involves an organized approach to identifying, analyzing and reporting on meaningful patterns of data from an interpretive perspective.

The analysis process had the following stages:

Phase 1: Familiarization with Data

All of the interviews were transcribed and listened to several times in order to become familiar with the participants' answers. At this

time, initial observations and significant statements were written down.

In Phase 2, initial codes are generated.Phase 2: Generation of Initial Codes:

Transcripts were systematically analysed and initial coding of the relevant portions of the transcripts were done. Some of the initial codes that were used were:

- Autonomous decision risks
- Instruction manipulation
- Data exposure concerns
- Memory contamination
- Permission management
- Human oversight requirements

This phase involves developing themes.The 3rd phase is Theme Development.

Codes that were related to each other were collapsed into larger category codes. Four main themes emerged by continual comparison and refinement:

As the use of AI increases, the cyber attack surface of agentic AI is also growing.As AI usage grows, so too does the cyber attack surface of agentic AI.

Prompt Injection Attack on AI Reliability:Prompt Injection Attack on AI Reliability:

4.Excessive Trust in the Ego Centricity of the Internet

4. Issues in autonomous tool use and governance.

Phase 4: Reviewing Themes

The identified themes were checked on the interview data for accuracy, relevance and consistency. Themes were developed to reflect participants' experiences.

Phase 5: Defining and naming themes

Each theme is well described in terms of its central theme and connection to the research questions.

Phase 6: Reporting Findings

The final themes were presented using participants' quotes to show the way cybersecurity professionals personally felt and interpreted cybersecurity challenges stemming from Agentic AI.

The study is trustworthy and has been conducted appropriately.

Four criteria for setting the quality and credibility of the qualitative research were taken from the work of Lincoln and Guba (1985) and considered:

Credibility

The process of lengthy interaction with the data, multiple transcribing of the interviews, and careful analysis of the participants' responses ensured credibility.

Transferability

Participants, research context and methodological procedures were described in detail so that the reader can determine the research applicability for other contexts.

Dependability

A systematic research process was kept, from documenting coding decisions, theme development and analytical procedures.

Confirmability

The researcher's objectivity was ensured by basing interpretations on the words of the participants and not on his own interpretations.

3.8 Ethical Considerations

Research was conducted in accordance with ethical principles. Participants were given information about the purpose of the study prior to participation. All participants gave informed consent and were told that participation was voluntary.

Participants' personal identity was not used, rather codes E1–E18 were given, thus ensuring confidentiality and anonymity. All the interview data were used only for research purposes, and participants could leave the study at any time.

3.9 Methodological Summary

The methodology used in this study allowed the researchers to delve deeper into the experiences of cybersecurity professionals in relation to cybersecurity challenges stemming from Agentic AI. The study employed a phenomenological approach and thematic analysis to gather the views of experts on the issues of prompt injection, memory poisoning, and autonomous tool use, offering insights into key risks for creating secure AI systems.

4. Findings

This study aimed to gain insight into the real-world experiences and perspectives of cybersecurity practitioners on the security challenges of Agentic Artificial Intelligence systems, with a specific emphasis on prompt injection, memory poisoning, and associated risks related to autonomous tool-use. Thematic analysis technique proposed by Braun and Clarke (2006) was used to analyze the data obtained from semi-structured interviews with 18 experts.

Based on the analysis four key themes emerged, reflecting participants experiences and interpretations:

2. Agentic AI is becoming a growing cybersecurity attack surface.2. Agentic AI is a cybersecurity attack surface that is growing.

Prompt injection is a major issue and a critical threat to the reliability of AI agents.Prompt injection is a key problem and critical threat to the reliability of AI agents.

3. Memory Poisoning as a Threat to AI Trust and Long-Term Security is a challenge from the memory.3. Memory Poisoning as a Challenge to AI Trust and Long-Term Security is a challenge from the memory.

The lack of autonomy in the use of tools is a risk in governance and control.

Themes, sub-themes and descriptions are provided in Table 2.

Table 2
Themes and Sub-Themes Generated from Interview Analysis

Main Themes	Sub-Themes	Description
Theme 1: Agentic AI as an Expanding Cybersecurity Attack Surface	Increased AI autonomy	Autonomous decision-making introduces unpredictable security risks
	Complex AI architecture	Integration of models, memory, and tools increases vulnerabilities
	Difficulty in monitoring AI behavior	Dynamic AI actions make traditional monitoring insufficient
Theme 2: Prompt Injection as a Critical Threat	Instruction manipulation	Attackers influence AI behavior through malicious prompts
	System information leakage	AI agents may reveal confidential instructions or data
	Difficulty in prevention	Natural language interaction creates security challenges
Theme 3: Memory Poisoning and AI Trust Issues	Persistent misinformation	Malicious information influences future AI decisions
	Memory validation challenges	Difficulty distinguishing legitimate and harmful information
	Long-term reliability concerns	Poisoned memory affects continuous AI operation
Theme 4: Autonomous Tool Usage and Governance Risks	Excessive permissions	AI agents may access unnecessary resources
	Unauthorized actions	Agents may perform unintended operations
	Need for human oversight	Human control remains essential for AI security

Figure 1
Thematic Map of Agentic AI Security Challenges

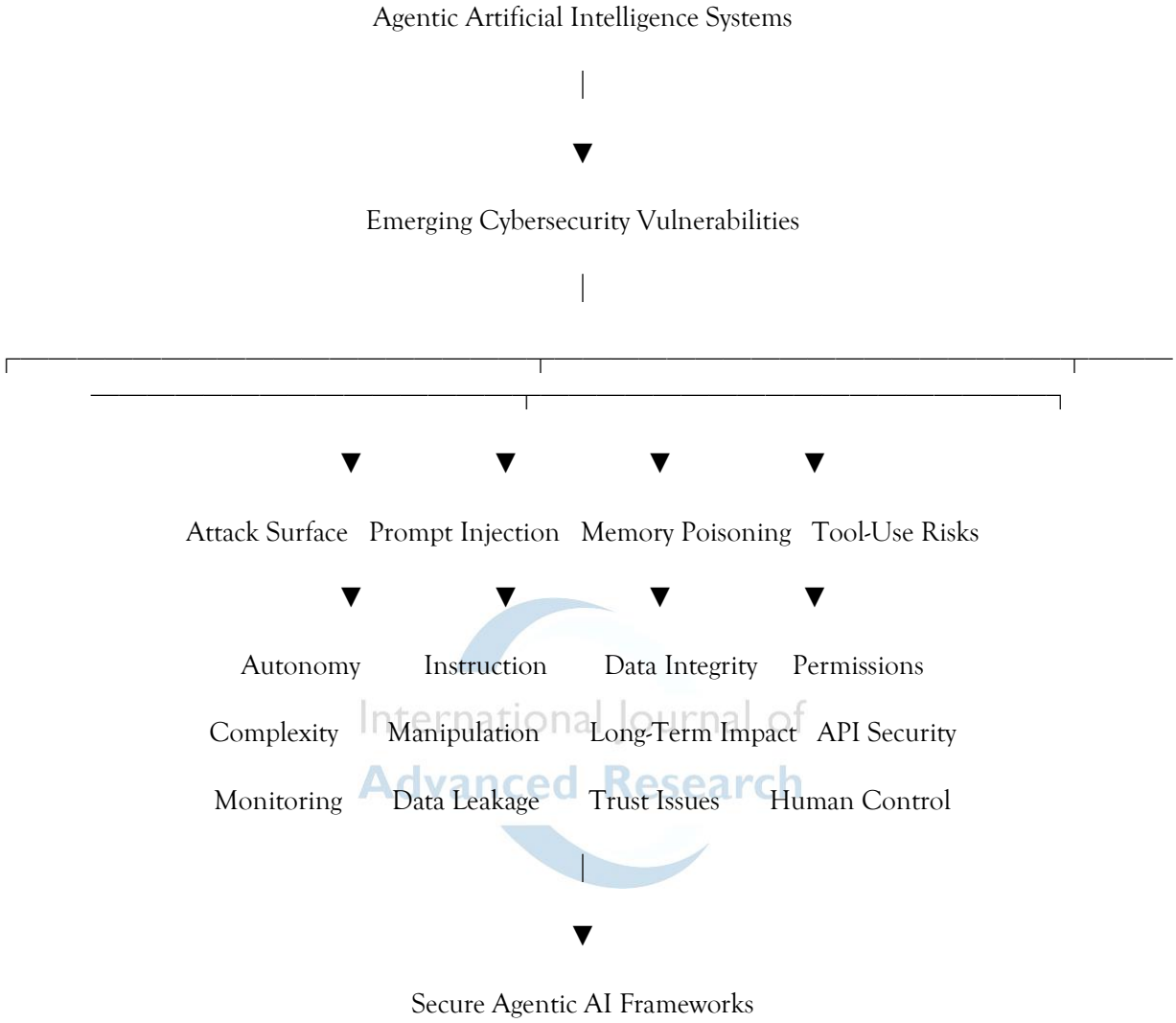

Figure 1. Conceptual representation of security challenges identified from expert experiences

Theme 1: Agentic AI as an Expanding Cybersecurity Attack Surface

The first big theme that emerged from the interviews with participants was the Agentic AI systems create a much larger cybersecurity attack surface than traditional software. The participants highlighted elements of autonomy, reasoning, memory and integration of external tools as sources of multiple security failures.

AI agents continuously understand information and act on it as it relates to the context, rather than following a pre-programmed set of instructions like applications do. Participants shared their insights on how this dynamic behavior complicates the prediction and control of AI actions.

One cybersecurity professional said:

The key distinction between traditional applications and AI agents is that the system is not just following commands; it's also understanding and determining how to carry them out. (E4)

Another participant highlighted:
"With access to a database, APIs, and external tools, each connection can be a security vulnerability." (E10)
Experts see AI autonomy as a new cybersecurity threat itself in these answers. While AI agents have the potential to plan and execute tasks independently to boost operational efficiency, this also brings potential vulnerabilities.
Sub-theme 1.1: **More autonomous and erratic behaviour by AI.**
Issues of security related to autonomy in decision making were often raised. AI agents can generate unpredictable results due to the intricate reasoning behind their actions, which are not programmed, they explained.
A participant explained:
"Traditional software will have the same predictable behavior, but AI agents can have different behavior based on context, memory and instructions." (E7)
This means that the uncertainty of the situation is a significant issue for security practitioners. AI systems must be autonomous, but security mechanisms need to be able to assess access to the system as well as AI reasoning processes.
The architecture of AI agents is complex.AI Agent Architecture is complex (sub-theme 1.2).
The presence of complexity was another major challenge identified by the participants. These are just a few examples of the many parts that modern AI agents can have:

- Large language models
- Memory systems
- Retrieval mechanisms
- External APIs
- Automated decision modules

Participants reported that weaknesses in any one element of the system could have an impact on the system's overall security.
One expert stated:
Language models are not the only issue; the entire system surrounding the model poses security risks. (E12)
The result indicates that a comprehensive strategy need to be adopted to obtain Agentic AI, which should include more than just the model itself.
Sub-theme 1.3: Monitoring the behavior of AI is difficult.Sub-theme 1.3: Monitoring the behaviour of AI is difficult.
Participants also pointed out that there are still difficulties in the monitoring of autonomous AI behavior. AI agents can respond differently when presented with similar situations, whereas traditional security monitoring methods are typically geared toward predictable activities of the system.
An AI security researcher said:
"It's important to continuously monitor because there are no guarantees that an AI agent will act identically in all situations." (E1)
The experts suggested better monitoring systems for recording the actions of AI, their tools, and any unusual behavior.
Theme 2: Prompt Injection: A Major Threat to AI Agent Reliability.Theme 2: Prompt Injection: A Critical Threat to AI Agent Reliability.
The second key theme was on prompt injection attacks. Nearly every participant pointed to the timely injection as one of the most critical dangers to Agentic AI systems.
Participants stated that AI agents heavily depend on natural language instructions, and attackers can take advantage of communication channels to make systems behave in a certain way.

Sub-theme 2.1: Instruction Manipulation
Prompt injection is defined by participants as a method of instruction manipulation and the use of an attacker trying to bypass the normal commands, they said.
One participant explained:
This means that there are situations in which the AI agent is not always aware of the instruction to follow, and this provides opportunities for attackers to influence its behavior. (E6)
Experts said there is still a huge amount of uncertainty about teaching the hierarchy of instructions in AI security.

Sub-theme 2.2: Confidential Information Leakage
Participants felt that if a prompt injection attack is successful, it could lead to the exposure of sensitive information.

A Cybersecurity consultant said:
"A manipulated instruction could force an agent to share information that is supposed to be confidential." (E5)
This sentiment was especially significant when agents can be granted access to private organizational information, as with enterprise applications of AI.

Sub-theme 2.3: Struggling with the complete prevention
The participants agreed that it is very difficult to prevent the occurrence of prompt injection completely as natural language is flexible.
One expert noted:
"Known attacks can be filtered, but attackers constantly are coming up with new manipulation techniques of instructions." (E14)
The study emphasizes the importance of multi-layered security solutions over depending on one security solution.
Memory Poisoning as a Challenge to AI Trust and Long-Term Security: Theme 3 of the AI Safety Summit.
The third theme discussed issues related to AI memory systems. Participants shared that although the use of memory helps to personalise and work efficiently, it also opens the door for ongoing attacks.

Sub-theme 3.1: Persistent Misinformation
The participants emphasized the importance of recognizing how harmful content in AI systems can impact future decisions.
One participant explained:
"A poisoned memory is more harmful than one poor response—they can impact all future interactions." (E11)
Memory poisoning was viewed as a long-term security problem as the malicious data could not be detected.

Sub-theme 3.2: Difficulty in Memory Validation
One of the big challenges identified by the participants was verification of stored information.
A researcher stated:
The system must determine if data going into memory is reliable or a threat. (E8)
There are several proposals for how AI agents can validate information sources prior to the addition of knowledge.There are a number of suggestions about how AI agents can validate information sources before adding information to their knowledge base.
This sub-theme explores the impact on AI Reliability.This sub-theme investigates the impact on AI Reliability.
The participants highlighted how flawed memory might affect the trust of users in AI systems.
One expert explained:
Trust is the basis of AI adoption: "If users don't trust an agent's memory, then they won't trust the AI to do a critical task." (E15)
Theme 4: Autonomy in using tools as a governance and control risk factor
The fourth theme was on the risks of AI agents acting autonomously on external tools and applications.
Participants agreed that the use of tools enhances the capabilities of artificial intelligence (AI), but also brought up significant security risks.

Sub-theme 4.1: Excessive Permissions
Unnecessary system access was cited as a significant risk by participants.
A security engineer said:
Unrestricted access of an AI agent is like to granting admin access to an unknown user. (E16)
Least-privilege access controls were highlighted as important.
Sub-theme 4.2: Unauthorized Autonomous Actions
There was concern that AI agents could carry out actions which are unintended with manipulation or incorrect configuration.
One participant stated:
The challenge is more than just keeping the attacks at bay, but also making sure the agent doesn't make bad decisions on his own. (E3)
Sub-theme 4.3: Importance of Human Oversight
There was a clear consensus that humans are essential in important AI processes.
A participant explained:

"Autonomous does not equal uncontrolled" (E18) – human supervision should be part of the security design.

The finding indicates that the experts believe that human-in-the-loop methods are crucial for responsible AI use.

Table 3

The key findings are summarized below:

KEY FINDINGS

RQ2: Monitoring issues of Agentic AI Monitoring problems, increased attacks surface, autonomy risks

RQ2: Prompt injection risks Instruction manipulation, data leakage, difficult prevention

RQ4: Memory poisoning impact Ongoing misinformation, trusting issues, reliability issues

RQ5: Areas of vulnerability in the security of public tools and equipment Access control, monitoring, human oversight

To give a general interpretation of findings.To provide an overall interpretation of findings

The findings reveal that cybersecurity practitioners consider Agentic AI security to be a multi-dimensional problem, related to technical, organizational and governance aspects. The need for specific AI-specific cybersecurity frameworks to secure AI agents was highlighted by participants.

The results indicate that while Agentic AI provides significant opportunities for automation and intelligent decision-making, its security depends on effective management of autonomy, information integrity, access privileges, and human supervision.

5. DISCUSSION

This study aimed to investigate the insights and attitudes of cybersecurity practitioners on the security issues surrounding AI systems such as Agentic AI, with a particular emphasis on prompt injection, memory poisoning, and the risk of autonomous tool-use. The study uncovered four key themes: (1) Agentic AI is becoming an expanding attack surface for cyber security, (2) prompt injection poses a significant risk of compromising AI reliability, (3) memory poisoning is a threat to AI trust and long-term security, and (4) the use of autonomous tools as a governance and control risk.

The results reveal that Agentic AI systems offer remarkable potential for automation, efficiency, and intelligent decision-making, but their autonomous nature poses intricate security challenges that aren't adequately addressed by traditional cybersecurity strategies. The findings are compared to the existing studies and relevant AI security frameworks.

5.2 Agentic AI is an Emerging Cybersecurity Attack Vector

The first big discovery was that Agentic AI systems are viewed as a new type of security problem for cybersecurity professionals, whereas traditional software programs are not. The participants noted that AI agents create more vulnerabilities, as they integrate reasoning, memory, external tools, and self decision-making. This discovery corroborates the previous studies published by Xi et al. (2023) on the greater AI attack surface that comes with transitioning from generative AI systems to autonomous agents. In the traditional cyber security paradigm, software behavior is predictable since programs follow a known sequence of instructions. But Agentic AI systems work through dynamic interactions between models, users, data sources and the external environment, which complicates security assessment.

The participants' concerns over the unpredictable behavior of AI further confirm the arguments of current discussion about the need to change the approach to AI security from traditional vulnerability management to continuous risk assessment. The NIST AI Risk Management Framework highlights the need for continuous monitoring of AI systems, as risks can arise during the AI lifecycle—from development to deployment to operational use—emphasizing the importance of ongoing monitoring of AI systems throughout the lifecycle (NIST, 2023).

The results also reveal that Agentic AI architectures are complex, making them challenging for security professionals. Unlike traditional applications, in which a vulnerability typically comes from a certain coding mistake, AI

agent vulnerabilities can arise from interactions between several components. A secure language model can be compromised if it's linked to an insecure database, poorly managed API, or external source of information.

This discovery aligns with the body of work that argues that AI security should consider a more systemic approach than just focusing on model-level security. Security strategies must take into account the full AI ecosystem, from data pipelines to memory systems, from tool integrations to human interactions.

5.2 **Prompt Injection is a major risk to AI reliability.Prompt Injection is a significant threat to AI reliability** (5.2).

The second key discovery revealed prompt injection as one of the biggest security threats of Agentic AI applications. The participants described how attackers can use natural language instructions to trigger inappropriate actions, circumvent security systems, or control the AI.

The discovery is in line with past research suggesting that the swift injection is a particular vulnerability of LLM-based systems since they process instructions not with explicit programming directives, but through language (Greshake et al., 2023). Unlike other injection attacks, like SQL injection, prompt injection doesn't take advantage of a certain software vulnerability. It does, however, take advantage of the flexibility and ambiguity of human language processing.

The concerns of the participants when it comes to instruction manipulation point to a core problem in AI security: the lack of secure definitions of instruction trust and distrust. Agentic AI systems often use information from various sources, such as external databases, websites, documents, and users. It's still a big challenge to decide what information should affect the AI's actions.

A previous study focused on the dangers of prompt injection, especially with the advent of AI system autonomy. While a simple chatbot might generate misinformation with a manipulated response, an AI agent that has access to external tools could result in unauthorized access to data, financial transactions, or changes to systems.

The results indicate organizations need to adopt more complex security strategies than simply relying on one security mechanism. Potential approaches include:

- Input validation mechanisms
- Instruction hierarchy controls
- Output monitoring
- Secure tool permissions

Approval from humans for sensitive actions

Prompt injection is highlighted as a significant vulnerability in the OWASP Top 10 for Large Language Model Applications, emphasizing the need for organizations to pay attention to this issue.

5.3 **Memory Poisoning and the Struggle for AI Trust over Time**

The third theme emphasized the issue of memory poisoning and its effects on the reliability of AI. The participants discussed how memory enhances personalization and efficiency while also creating potential new security threats due to the potential for malicious information to remain stored and impact future AI judgments.

This discovery is further research on AI data security, showing that integrity has come to be a necessary data security requirement for autonomous systems. Traditional data management practices can be used to fix faulty data in a conventional software application stored in a database. AI memory systems add complexity, though, due to the way that information stored is used in reasoning and affects responses in "the future.

The concerns raised by the participants about the continued misinformation align with the findings of past research on the limitations of retrieval-augmented generation (RAG). In the era of AI agents that access external data sources, malicious data inputs can have a long-term impact on the quality of AI results.

One important takeaway from this discovery is that the field of AI security has progressed from merely safeguarding integrity and availability to ensuring information integrity. A flawed memory can make an AI system continue to function, but make poor decisions.

The participants highlighted the need for mechanisms to ensure memory validation. These may include:

Preverification of source before storing information.Source verification prior to storage of information.

Retrieved knowledge is trusted according to the score

- Periodic memory review
- Protect memory access restrictions
- Abnormal information patterns detected

These strategies are especially crucial for enterprise AI implementations where misjudgments can have substantial operational repercussions.

5.4 Autonomous use of tools and governance issues

The fourth key insight was that the ability to use tools independently is the most powerful asset and a security threat of Agentic AI systems. Participants acknowledged the benefits of integrating AI agents with external tools, APIs, and databases for robust automation, as well as the potential for illicit usage.

This discovery aligns with earlier studies that highlight how AI agents with too many privileges can lead to security risks akin to those of hacked human accounts. Unlike people, AI agents can also perform actions in a quick manner and across several systems, which means that security breaches can have a greater effect.

The principle of least privilege was mentioned countless times among the participants. AI agents should only be granted access to resources that are needed for the specific tasks they will be performing. More permissions mean more potential for misuse, whether it be due to malicious attack, system error, or flawed AI reasoning.

The results also highlight the need for human supervision. While Agentic AI systems are intended to be autonomous, the participants noted that the lack of oversight leads to unacceptable risks, especially in sensitive areas.

This aligns with the concept of "human-in-the-loop" AI governance, where AI systems carry out automated activities but vital decisions are still scrutinized by humans. This is a method that can be efficient and yet held accountable.

5.5 Theoretical Implications

The results are valuable additions to the AI security body of knowledge in a number of ways. The first thing it does is it brings to light a deeper understanding of AI security beyond just technical vulnerability analysis, as it looks at the experiences of professionals.

Second, the results underscore that there are organizational and ethical and governance aspects to Agentic AI security, in addition to technological aspects. The collaboration among AI developers, cybersecurity experts, policymakers, and end users is essential for effective security.

Thirdly, the research shows that autonomy introduces a new cybersecurity risk dimension. Traditional cybersecurity frameworks mainly focus on protecting systems from external attacks; however, Agentic AI requires consideration of risks emerging from the AI system's own decision-making processes.

5.6 **Practical Implications**

The results have several implications for companies that use Agentic AI systems.

For AI Developers:

Developers should incorporate security principles throughout the AI development lifecycle. Prompt injection testing, memory integrity testing, and tool-permission testing should be part of the security testing.

For Organizations:

Organizations need to create AI governance structures to define:

The use of AI with acceptable policies.

- Access control requirements
- Monitoring procedures
- Incident response strategies

For Cybersecurity Professionals:

Security teams need to build up unique knowledge of vulnerabilities around AI and implement continuous monitoring strategies for autonomous systems.

For Policymakers:

Regulatory frameworks needs to reflect the particular risks of autonomous AI systems, and foster responsible innovation.

5.7 Contribution of the Study

This research adds to the growing Agentic AI security body of work by offering empirical information on cybersecurity professionals' understanding and reaction to new threats introduced by AI. Most previous work focused on technical vulnerabilities, but here the focus is on the experiences and concerns of professionals tasked with securing intelligent systems in practice.

Overall, the findings highlight the need for a holistic strategy to achieve Agentic AI, which must address technical, governance, and human oversight considerations. This research highlights the critical issues of prompt injection, memory poisoning, and the unauthorized use of tools, offering guidance for future developments in AI security policies.

5.8 Discussion Summary

In summary, the results reveals that Agentic AI systems are a promising and a difficult challenge to cyber security. They are capable of being highly autonomous, which is great for applications, but also means new capabilities that can be exploited in different ways and need special security solutions.

The study underscores the need for a new approach to AI security, beyond what current cybersecurity solutions can achieve. Instead, adaptive security frameworks should be implemented that are designed for the threats posed by AI, are human-centric, and for which autonomous operations can be trusted.

I will finish the rest of the sections and include the APA 7th edition reference list with verified DOI links to the sections. I will not include any unverified DOI numbers, references with verified DOI numbers should be added for submission to HEC Y/ESCI.

6. CONCLUSION

Agentic Artificial Intelligence is a major paradigm shift in the evolution of AI systems, as it introduces autonomous reasoning, planning, memory usage, and interaction with external tools. These capabilities can be incredibly useful for automation, productivity, and decision-making, but they also raise some interesting cybersecurity issues that necessitate new strategies for how to protect AI systems.

In this research, we examined the insights of cybersecurity practitioners about the security threats of Agentic AI systems, such as prompt injection, memory poisoning, and autonomous use risks. The study used a qualitative phenomenological approach that showed that experts see Agentic AI as a rapidly growing field of cybersecurity in which current security mechanisms may not be adequate.

These results confirmed that AI agents introduce more attack surface due to their complexity, external system integrations and autonomous decision making. Prompt injection was recognized as a major risk by participants, as it could lead to the injection of malicious instructions into AI models, potentially compromising their reliability and functionality. Likewise, memory poisoning was identified as a major concern, as tainted data could impact subsequent AI decisions and undermine user trust. Moreover, the use of tools without any human supervision was highlighted as a key governance issue, within which a variety of risks were found, such as granting too many permissions, too many actions, and not enough human control.

The study finds that a holistic strategy is needed to ensure Agentic AI is secure, including technical measures, governance protocols, and ongoing human monitoring and oversight. Traditional cybersecurity approaches must be extended to build protection mechanisms for AI itself: model protection, memory integrity, instruction security, and management of access to AI tools.

With the adoption of Agentic AI growing across industries, it is important to ensure its operation in a trustworthy and secure manner. The research work presented here helps to fill the growing body of literature on AI security by giving expert perspectives on real-world challenges and

solutions to creating more secure intelligent autonomous systems.

7. **Recommendations**

Based on the above study, the following recommendations are made:

7.1 **It is about the creation of AI-specific security frameworks.In 7.1, it is about the development of security frameworks specific to AI**.

An organization should establish specific policies and procedures for Agentic AI systems. These frameworks should include consideration of:

- Prompt injection protection
- Memory validation mechanisms
- Secure tool integration
- AI behavioral monitoring for continuous data collection
- Risk assessment procedures

Current cybersecurity approaches need to be modified to account for AI-related vulnerabilities.

7.2 Implementing Least-Privilege Access Control

Access to permissions for AI agents should be limited to what is required to accomplish assigned tasks. Don't give anyone free access to the database, APIs, or operational systems.

There are several elements to access control mechanisms that should be included:

- Permission limitations
- Authentication procedures
- Activity monitoring
- Automated access reviews

To monitor the behavior of AI continuously.To monitor the behavior of AI continuously.

Organizations must have systems to monitor Agentic AI to detect:

- Abnormal AI behavior
- Suspicious tool usage
- Unauthorized decisions
- Potential security breaches

Real-time monitoring can minimize the effects of AI security incident impacts.

To enhance the security of AI memory.To improve the security of the AI memory.

AI memory enabled organisations should have memory protection policies such as:

Verification of information stored on disk.

- Trusted data sources
- Memory auditing

The ability to identify harmful content.Ability to identify harmful content.

To guarantee the reliability of AI decision making, memory integrity is crucial.

7.5 **Maintaining Human Oversight**

While AI agents are supposed to be autonomous, there should be human oversight in critical decisions. Human in the Loop methods can help avoid negative consequences and ensure greater accountability.

7.6 **AI Security Training**

There should be training programmes for:

- Cybersecurity professionals
- AI developers
- System administrators
- End users

The emphasis needs to be on understanding threats associated with AI and on adopting responsible practices for securing AI.The training should be targeted to understanding the threats associated with AI and adopting responsible practices to secure AI.

The study has its limitations.

Although useful information was gained, there were a number of limitations to this study.

The study was carried out with 18 participants using a qualitative phenomenological design, first. While the results were useful for the study of expert experiences, it ought not to be assumed that it reflects the experiences of all cybersecurity experts around the world.

Second, the participants were chosen using purposive sampling and this could restrict the variety of views.

Thirdly, the study used self reported experiences and professional opinions of the participants. Actual technical testing of AI systems was beyond the scope of this research.

Finally, Agentic AI technologies are constantly changing, and new security threats are likely to emerge in the future that are not included in this study.

9. Future Research Directions

This study could be continued in the future by:

9.1 Quantitative Investigation

Future research can investigate the connections among AI awareness, readiness, and implementation of security practices.

9.2 Technical Vulnerability Testing

Experimental studies can be performed by researchers to test the prompt injection, memory poisoning and tool-use attacks in the real world.

9.3 Cross-Country Comparative Research

International comparison studies may be conducted to examine differences in the security practices and regulations of AI across countries.

9.4 development of AI security models

Future studies should aim to develop comprehensive security structure specifically tailored to autonomous AI agents.

9.5 Organizational Case Studies

Practical value could be gained from detailed case studies involving the implementation of Agentic AI systems within organizations that could offer insights into security management practices.

REFERENCES

Greshake, K., Abdelnabi, S., Mishra, S., Endres, M., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. https://doi.org/10.1145/3605764.3623985

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* National Institute of Standards and Technology.https://doi.org/10.6028/NIST.AI.100-1

OWASP. (2025). *OWASP Top 10 for Large Language Model Applications.*
https://owasp.org/www-project-top-10-for-large-language-model-applications/

Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models.https://arxiv.org/abs/2211.09527

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., & Zhang, W. (2023). The rise and potential of large language model based agents: A survey.https://doi.org/10.48550/arXiv.2309.07864