University of Kent

Kent Academic Repository

Gupta, Shivangi, Arief, Budi and de Lemos, Rogério (2026) *Agentic AI vs Non-Agentic AI: Motivation, Security Implications, and Research Foundations*. In: 2026 21st European Dependable Computing Conference Companion Proceedings (EDCC-C). . IEEE CPS ISBN 978-1-971299-06-8.

Downloaded from
https://kar.kent.ac.uk/114337/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.23919/EDCCCCPS00002.2026.00041

This document version
Author's Accepted Manuscript

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information
For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Agentic AI vs Non-Agentic AI: Motivation, Security Implications, and Research Foundations

Shivangi Gupta, Budi Arief, and Rogério de Lemos
School of Computing, University of Kent, Canterbury, United Kingdom
Email: {sg106, b.arief, r.delemos}@kent.ac.uk

***Abstract*—A significant change in the development and application of artificial intelligence (AI) systems is the transition from non-agentic to agentic AI. Non-agentic AI systems, such as prompt-based language models and classical machine learning models, operate reactively, producing outputs only in response to inputs. They lack long-term memory, long-term goals, or the capacity to act independently. Agentic AI systems, on the other hand, are made to act more autonomously through goal-setting, multi-step planning, tool use, memory storage, and action execution in physical or virtual environments. This paper aims to explain the emergence of agentic AI, distinguish it from non-agentic AI, and examine the new security and governance challenges arising from this novel mode of operation. The approach used in this paper includes research about intelligent agents, large language models (LLMs) based agents, AI security, and governance frameworks. The paper also highlights how autonomous behaviour increases AI attack surface, shifts security concerns, and focuses from isolated model errors to risks involving decision-making logic, persistent memory, delegated permissions, and long-running agent behaviour. Finally, the paper argues that although agentic AI may increase the threat surface, it can be deployed responsibly provided that appropriate system-level safeguards are in place. This highlights the need for new or extended security and authorisation frameworks focusing specifically on agentic AI.**

***Index Terms*—Agentic AI, Autonomous systems, Trusted Delegation, AI safety, ML models.**

I. Introduction

Originally, AI was developed to support human decision making rather than to autonomously replicate or replace it [1]. Traditional AI systems, including rule-based expert systems, decision trees, and supervised learning models, operate under strict human supervision and perform predefined tasks within well-defined contexts [1], [2]. Although statistical learning and deep neural networks matured, their deployment remained unchanged, namely, non–agentic [3]. The models receive inputs and generate outputs, leading to execution getting terminated without maintaining persistent goals, memory, or control over external activities [3]. The large language models (LLMs) currently in use exhibit advanced reasoning and generative capabilities. However, when deployed as chatbots or decision support tools, they remain fundamentally reactive: they do not initiate actions, pursue independent goals, or interact with external systems unless explicitly instructed to do so [4], [5].

The growing complexity of real-world digital environments has exposed fundamental limitations of non-agentic AI. Domains such as enterprise operations, cybersecurity defence, cloud orchestration, scientific discovery, and cyber-physical systems increasingly require advanced reasoning, multi-step task execution, and coordination across heterogeneous tools and services [6], [7]. These operational requirements led to the emergence of *agentic AI*.

Agentic AI systems differ from conventional AI systems in their ability to formulate goals, plan multi-step actions, maintain persistent memory, and execute actions in external environments through tools, application programming interfaces (APIs), or actuators [8]. The architecture of these systems comprises reasoning engines, orchestration layers, memory subsystems, and execution interfaces, enabling them to operate continuously and adaptively within dynamic environments [9]. This development represents a fundamental shift in AI capability: rather than optimising isolated predictions, such systems manage decision-making processes over time. As a result, agentic AI is more in line with the ideas of intelligent agents in autonomous planning and multi-agent systems (MAS) [8].

While autonomy opens the door for new, powerful applications, it also changes the AI security and risk landscape. Non-agentic AI security research has traditionally focused on data- and model-centric threats, such as data poisoning, model extraction, and privacy leakage [10], [11]. Agentic AI introduces behaviour-centric threats that arise from the system's ability to act, remember, and pursue goals over time. These include prompt injection attacks that manipulate agents' goals, memory poisoning that alters future behaviour, unauthorised or unsafe action execution via tools, and cascading failures in multi-agent environments [12], [13], [14].

Seen in this light, reinforcement learning (RL) helps explain why agentic AI emerged but also why RL cannot, by itself, address the demands of modern autonomous systems [15]. RL demonstrated that goal-directed behaviour over time is both feasible and valuable, yet its reliance on closed environments, fixed reward specifications, and offline optimisation exposed practical limits when systems are deployed in open, evolving digital infrastructures [16].

More recent work on deep RL, hierarchical RL, and multi-agent RL has moved closer to agentic characteristics by enabling longer-horizon planning, partial observability handling, and decentralised coordination [15], [16]. Nevertheless, these systems still differ fundamentally from contemporary agentic AI architectures. While agentic AI systems pursue dynamically changing goals, use symbolic reasoning, preserve long-term memory, and dynamically call external services via APIs and tools, RL agents, on the other hand, optimise a

single, explicitly stated reward signal [15], [17]. In this way, agentic AI can be viewed as a hybrid system that integrates deliberative planning, language-based reasoning, and system-level orchestration with elements of decision-theoretic control from RL [15].

Agentic AI offers compelling advantages in scalability, adaptability, and resilience that are increasingly indispensable in modern digital infrastructures. Continuous monitoring, autonomous remediation, adaptive decision-making, and workflow orchestration are difficult to achieve on a scale through human oversight alone [18], [19]. As such, the central research challenge is not whether agentic AI should be adopted, but how autonomy can be constrained, governed, and secured without undermining its benefits. This paper addresses this challenge by systematically analysing the motivations behind agentic AI, contrasting it with non-agentic systems, examining the unique security risks introduced by autonomy, and synthesising emerging defence mechanisms and research gaps. By doing so, it lays a theoretical foundation for future work on secure, trustworthy, and governable agentic AI systems.

Contributions. This paper provides four contributions:

- Analysis of different security risks and defence challenges faced in agentic AI, such as prompt injection, memory poisoning, and privilege escalation;
- Evaluation of various defence strategies and mitigation approaches required for agentic AI systems and why existing approaches are insufficient for such systems;
- Evaluation of the challenges associated with integrating security mitigation strategies into AI agentic systems;
- Identification of research gaps and highlighting the need for enhanced versions of existing mechanisms, giving rise to new research directions.

II. Background and Conceptual Foundations

A. Non-agentic AI Systems

Non-agentic AI systems are defined as stateless systems having reactive behaviour [4]. They generate outputs based on the inputs, without maintaining persistent goals or executing autonomous actions. Some examples include supervised learning models used for classification, recommendation engines, and LLMs deployed in advisory roles without task execution capabilities [10], [14]. These systems produce deterministic outputs and operate under human supervision without affecting the environment beyond pre-defined instructions [4].

Non-agentic AI has well-researched security risks. Data privacy breaches can occur via membership inference or model inversion attacks [13]. Adversarial inputs can be crafted to mislead models [12]. Model extraction attacks may allow adversaries to reconstruct proprietary models through API queries [20]. Since these systems do not behave autonomously, associated threats can often be mitigated through procedural oversight [18]

In real-world scenarios, non-agentic AI systems remain valuable when requirements include human-supervised, sequential decision-making tasks, such as spam filtering and customer support chatbots [14].

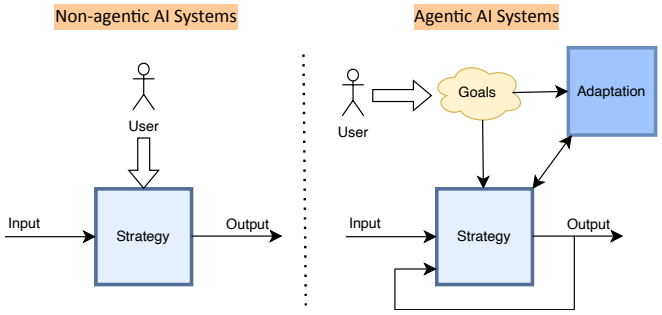


Fig. 1. Reactive non-agentic AI versus Goal-driven agentic AI

B. Agentic AI Systems

Agentic AI extends the notion of non-agentic systems: agentic AI systems behave beyond reactive computation. They are autonomous systems integrating perception, reasoning, planning, and action execution [3]. Unlike non-agentic systems, agentic AI systems create multi-step plans and retain memory to inform future decisions [3], [19]. Architecturally, they combine reasoning engines, orchestration layers, memory stores, and tool interfaces [19].

Autonomous execution facilitates interactions with software systems, APIs, or physical actuators [19], which contrasts with how non-agentic systems operate. But they also expand the threat surface, thus requiring runtime governance, behaviour monitoring, and identity management [1], [19].

Figure 1 shows an operational diagram that depicts the key distinction between agentic AI (on the right) and non-agentic AI (on the left). Agentic AI systems are goal-driven. Agentic systems begin with a goal and leverage their strategy-building capabilities to decompose it into subtasks. They then execute these tasks using available tools and APIs, observing the outcomes and storing them in memory. This feedback allows the agents to learn and refine their strategies, improving efficiency in achieving their goals [4], [19]. In agentic AI systems, strategy is determined by their goals, inputs, and adaptation process

In contrast, non-agentic AI systems exhibit linear, reactive behaviour: they process given inputs to produce corresponding outputs without altering their strategy, and execution terminates thereafter. Such systems are stateless, lacking persistent memory and any capacity for autonomous operation.Every interaction is independent and requires human intervention to continue [4], [19]. In non-agentic AI, strategy is decided based on the inputs given by humans.

Table I grades an AI system's agentic strength from "Basic" to "Fully Autonomous" [29], [30], [26]. Suppose our goal is to book a flight In *Low Agenticness* (Basic to Moderate Autonomous), an agent finds flights, but the human must enter details into the website and pay. In *High Agenticness* (Moderate Autonomous to Fully Autonomous), an agent looks up the calendar, finds the best flight, calls the booking API, pays with a virtual credit card, and updates the calendar with confirmation, adapting if the first flight is sold out.

TABLE I
RUBRIC FOR AGENTICNESS

Dimensions	Basic	Moderate Autonomous	Fully Autonomous
Autonomy [21], [22]	Requires constant human intervention	Operates on its own for most of the task but requests human input only during blockers	Fully independent and requires human oversight only for critical final approvals
Goal Orientation [23]	Responds to one interaction prompt	Sets a primary goal and breaks it down into 2-3 logical sub-tasks	Based on the goal, it breaks the goal into sub-tasks and re-prioritises tasks based on evolving context
Reasoning and Planning [24], [25]	Basic reasoning	Uses chain-of-thought to plan steps and has basic self-correction	Develops complex, multi-step strategies; learns from failures to adapt future plans
Tool Use [25], [26]	No tool integration	Calls APIs/tools to retrieve data or perform basic tasks (e.g., search, email)	Dynamically selects, uses, and chains multiple tools (e.g., browsing, coding) without prompt guidance
Memory [25], [27]	Zero memory and every turn is new	Maintains short term memory	Uses long term memory to save history across sessions
Safety and Guardrails [28]	Basic built-in safety controls	Uses hard-coded safety rules (e.g., "do not pay above £300")	Detects ethical, safety, or privacy issues and stops actions that might cause harm for review

C. Case Study

To explain the key differences between the two systems, we selected a scenario that demonstrates the capabilities of agentic AI systems that non-agentic AI systems lack [4], [19], [31], taking as an example an Enterprise Compliance and Regulatory Change Management system that was previously non-agentic [32], [33]. Suppose a bank operates in multiple countries, and each country has its own regulations. If the bank fails to comply with the regulations of some countries, that could lead to legal and financial issues for the company [34].

In a non-agentic system for this scenario, a compliance officer would get an update regarding a change in regulation [34]. This officer can ask the non-agentic system to summarise the change, the category of customers getting impacted, and what they are required to do. In response, the non-agentic system may reply with: a summary of the change, a list of existing category 'A' customers that would be impacted, and how such customers would need to re-verify their identity within 90 days using the Know Your Customer (KYC) feature [31], [35]. The system replies to the tasks one at a time, and stops after all output is produced, i.e., it does not do anything further, decide what needs to be done next, or start the re-verification process on its own. In this scenario, the human operator (i.e., the compliance officer) is required either to tell the system what to do next or to do it themselves. Although such a non-agentic system can support human operators, it cannot work on its own, since human intervention is required at all times.

In contrast, an agentic AI system based on the same scenario would start with creating sub-tasks, such as: (i) do continuous monitoring and detect any new change as soon as it comes (without any human intervention) (ii) create a list of impacted category 'A' customers and then initiate re-KYC for impacted customers; (iii) collect the KYC documents; (iv) send results with the list of impacted customers; (v) prevent deadline breach, etc. While executing, if any sub-task (which may impact the overall goal) is unachievable under the current strategy, agentic systems can autonomously adapt the strategy, unlike non-agentic systems, which rely on pre-defined rules/tuning (i.e. they require outside involvement).

Several articles address real-world implications of this case study, in which banks have updated their systems to agentic AI systems or are in the process of doing so. Verhagen et al [34] states that, according to Interpol, banks can detect only two per cent of financial crime flows, despite significant expenditure on the KYC process. Another article [35] talks about a financial service company that has upgraded its KYC process with agentic AI, and they were able to handle 3.5 times more cases per month, and processing time dropped to 2-4 hours from 5-7 days, with 99.2% of accuracy achieved in the document checking process.

III. RELATED WORK

Previous research in intelligent agents, autonomous planning, and multi-agent systems (MAS) provided the early foundation to agentic AI. Such research shows that agents are entities capable of perceiving their environment and acting on it to achieve their goals [2]. Long before LLMs were introduced, other models such as Belief–Desire–Intention (BDI) models laid the theoretical groundwork for autonomy [1], [2]. In previous studies, the primary focus has been on correctness, coordination, and efficiency, with security receiving comparatively less attention.

With the rise of machine learning, particularly deep learning, research attention shifted towards the security of non-agentic AI systems. Extensive literature exists on adversarial examples [3], [4], [5], [6], [7]. However, these works assume AI systems do not autonomously execute actions or maintain persistent goals. They assume systems to be: reactive, stateless, and non-autonomous. As a result, the proposed defences, such as adversarial training or differential privacy, are insufficient for agentic deployments.

Recent work also examines LLM agents and tool-augmented reasoning systems. Research on LLM agents shows us that they have the capabilities which can be useful in agentic AI systems [8]. Existing frameworks, such as *ReAct*, integrate reasoning with action execution, enabling agents to interleave decision making with tool use [9]. Similarly, *Voyager* demonstrates how large language models can support open-ended exploration and self-directed skill acquisition [36].

Literature also positions RL as a bridge between non-agentic AI and fully agentic AI [23]. Traditional RL has been applied to sequential decision-making tasks in which an agent interacts with a fixed environment and optimises cumulative

reward, as seen in early robotics and game-based domains [15]. However, such systems typically rely on externally defined reward signals and fixed state–action mappings, with limited context awareness or adaptive long-term goal formulation [15]. In the context of LLMs, RL has been used to fine-tune models for interactive behaviours [15], [16]. Recent surveys highlight that agentic RL reframes conventional RL by treating the environment as partially observable, enabling adaptive behaviour spanning planning, tool use, memory, and multi-task coordination [16].

Building on this perspective, current research increasingly integrates RL into larger agentic architectures that combine RL with reasoning, memory, and external tool interfaces. For example, recent empirical studies show that RL can enhance LLM agents' reasoning efficiency and tool-use performance, if provided with careful design of training and reward structures that account for long-horizon dependencies [15], [16].

Yet, even as RL algorithms evolve to support multi-turn, multi-task strategies in agentic RL frameworks [16], [23], the literature acknowledges that fully autonomous systems require hybrid designs, like LLM reasoning cores combined with RL policies, persistent memory, and policy optimisation mechanisms in open environments [23]. These hybrid systems reflect the transition from non-agentic reactive models towards architectural autonomy.

In response to emerging threats, several governance and risk-management frameworks have emerged. Industry and regulatory efforts such as NIST's *AI Risk Management Framework (AI RMF)* [12], and MITRE's *ATLAS* [37] provide structured taxonomies of AI risks and adversarial techniques [20]. Gartner's *Trust, Risk, and Security Management (TRiSM)* framework [38] focuses on organisational controls, monitoring, and compliance for deployments involving AI [18].

Other initiatives, such as *AGENTSAFE* [39] and work on secure agent orchestration, try to formalise control mechanisms for autonomous AI systems [18], [19]. These mechanisms introduce concepts such as runtime policy enforcement, scoped autonomy, and human-in-the-loop oversight. However, existing frameworks are high-level and descriptive, i.e., they lack formal threat models, proper metric evaluation and verified security guarantees associated with agentic AI [18], [19], [40].

As literature reveals, the research is fragmented. Even though agentic AI is becoming more popular, there are still several open research gaps: (i) There is a lack of formal threat models that focus on autonomous, goal-driven AI systems, since existing AI security frameworks largely talk about software security, taking non-agentic AI assumptions into account; (ii) Evaluation methodologies for agentic AI security are largely absent, making comparative analysis difficult; (iii) Explainability and accountability mechanisms for agentic decision-making remain underdeveloped, particularly in multi-agent settings where responsibility is distributed across interacting entities; (iv) Identity and access management models for autonomous agents are still immature; (v) Regulatory and ethical frameworks lag technological developments, leaving open questions about liability, compliance, and societal impact.

IV. Methodology

To address the research gaps mentioned above, we formulated the following research questions (RQs):

- RQ1: What are the security risks and mitigation strategies possible in agentic AI systems?
- RQ2: What are the challenges organisations are facing in integrating mitigation strategies into agentic AI systems?
- RQ3: How can we take advantage of agentic AI's capabilities with proper approaches if incorporated in real-world scenarios despite having security risks?

We selected papers that closely align with our research using the following terms: "agentic AI", "non-agentic AI", "autonomous systems", "reinforcement learning", "RL", "trusted delegation", "large language models", "LLMs", "stateful vs stateless", "reactive behaviour in non-agentic AI", "threat models for agentic AI", "security and mitigation approaches for agentic AI", "OAuth and trusted delegation", "defence challenges in agentic AI", "multi-agent systems", and "MAS".

For our research, we used IEEE Xplore, Science Direct, and ACM DL. We identified 60 relevant papers for our analysis, of which 40 were used in the research and helped inform our literature review. The remaining 20 papers were excluded using abstract filtering.

V. Results

Agentic AI systems introduce risks that are distinct from those of non-agentic AI systems. Unlike old attacks that target model outputs, agentic attacks may exploit the decision-making loop of the agent [4].

A. Major Security Risks Associated with Autonomous Systems

Table II highlights the major security risks affecting agentic and non-agentic AI systems and how the behaviour of these risks differs in both systems.

1) Unauthorised Action Execution: Like non-agentic AI systems, agentic AI systems can initiate sequences of actions across systems [3]. If an agent is compromised, malicious actors can manipulate it to execute unauthorised commands, potentially across internal and external infrastructure [8], [9]. For instance, like non-agentic AI, an agentic AI deployed in cloud orchestration could maliciously delete critical resources or reconfigure network policies, resulting in cascading failures [8]. However, the situation can be serious in agentic AI systems because they can store information in long-term memory and create sub-tasks based on the goal (unlike non-agentic systems). An example of this risk class is prompt injection attacks, in which malicious inputs can manipulate agent reasoning and decision pathways.

2) Privilege Escalation via Tool Misuse: In agentic AI systems, agents often interact with tool APIs and external services to perform multi-step tasks [9]. There is a possibility of broad permissions being granted to facilitate autonomy. If an attacker gains control of the agent, tool misuse may escalate privileges beyond intended scopes, enabling access to sensitive systems, credentials, etc. [9]. Traditional Role-Based Access Control (RBAC) is insufficient in this context,

TABLE II
COMPARISON OF SECURITY RISKS IN NON-AGENTIC AND AGENTIC AI SYSTEMS

Risk Category	Non-agentic AI	Agentic AI	Explanation
Unauthorised Act Execution (incl. Prompt Injection)	The attack can cause incorrect outputs or misclassification [3].	It manipulates long-term goals, decision-making, or planning loops [3].	In agentic AI systems, a single malicious prompt can propagate through planning and tool execution layers, causing failures or misaligned actions [3], [8], [9], [10].
Tool Misuse / Privilege Escalation	Not applicable since there is no autonomous action execution [19].	In agentic AI systems, an attack can cause unauthorised execution of system-level tools, APIs, or external services [19].	Autonomous agents with broad permissions can unintentionally escalate privileges or execute harmful operations without oversight [19], [20].
Memory Poisoning	Not applicable, as the systems are stateless [1].	The attack can lead to persistent manipulation of intermediate reasoning, contextual knowledge, or long-term preferences [3].	In agentic AI, attacks can affect future actions, e.g., create systematic misbehaviour, or affect multi-agent coordination [1], [3], [13], [41].
Goal Manipulation / Drift / Misalignment	Not possible because the goals are externally imposed and static [8], [42].	The attack can induce progressive divergence between the agent's intended goal and its actual behaviour due to environmental influences or malicious inputs [42].	The attack may spread over multiple steps, leading to unsafe behaviours in complex environments [8], [9], [42].
Complex Multi-Agent Vulnerabilities	Non-agentic systems are not affected due to the limitation to single output or immediate system impact [8].	Attacks can cause system-wide damage due to their multi-agent, cross-tool, or cross-domain effects [8].	Agentic AI attacks can affect multiple autonomous agents, thus increasing risk [8], [9], [14].
Identity, Authentication, and Lifecycle Risks	Non-agentic systems execute predefined logic and have no self-directed behaviour [12], [20], [43].	Agents in an agentic environment act as first-class digital identities with delegated entitlements [12], [20], [43].	In agentic systems, risks of misuse, escalation, and accountability gaps increase due to complex delegation, dynamic credentials, and context-specific authority [12], [20], [43].

as agents in agentic AI systems require fine-grained, context-aware permissions [12], [20].

3) Memory Poisoning and Goal Manipulation: In agentic AI systems, persistent memory plays a central role in storing contextual knowledge, intermediate reasoning steps, and long-term preferences [41]. While this supports long-term learning and adaptive behaviour, it also exposes the agent to memory poisoning attacks, in which malicious data alters reasoning paths [41]. Similarly, goal manipulation attacks exploit prompt injection vulnerabilities, leading to subtle modifications in the agent's goals subtly over time, potentially causing systematic drift from intended behaviour [42]. In MAS, a single compromised agent can propagate erroneous strategies across peers, amplifying systemic risk [42].

4) Complex Multi-Agent Vulnerabilities: When multiple agents operate in an environment where coordination is crucial, inter-agent dependencies increase the potential impact of attacks. For example, a compromised agent propagating false information could lead other agents to execute unsafe plans. These failures introduce risks that can lead to systemic failures in distributed control systems [14]. Existing security frameworks do not sufficiently address adversarial dynamics in multi-agent interactions.

5) Identity, Authentication, and Lifecycle Risks: Agentic AI poses new security risks to existing identity and access control mechanisms. Agents often operate with broad permissions to function effectively, but this violates the principle of least privilege [14]. Treating agents as first-class digital identities introduces unresolved challenges regarding authentication, authorisation, accountability, and lifecycle management [12], [20]. Unlike static user accounts, agents are short-lived, adaptive, and autonomous, and therefore require continuous integrity verification and policy constraint enforcement [19]. Failure to enforce proper identity management could allow unauthorised access, privilege abuse, or misattribution of agent actions [44].

B. Defence Strategies and Mitigation Approaches

Mitigating security risks in agentic AI requires solutions beyond traditional input-output security models [18]. Several key mitigation strategies are outlined below.

1) Runtime Governance: Runtime governance enforces policy-based constraints on agent actions in real time [11], [18]. Enforcement layers monitor all tool invocations, memory accesses, and goal-directed decisions, intercepting unsafe or non-compliant actions before they propagate. Runtime governance can include dynamic rule-checking, pre-commit validation, and automated rollback mechanisms, ensuring that autonomy operates within a secure envelope [12], [18], [42].

2) Behavioural Monitoring and Anomaly Detection: Modelling normal agent behaviour, such as action sequences, tool usage patterns, or trajectories in goal evolution, can help us proactively identify deviations indicative of compromise [19]. Advanced approaches include: (i) Sequence anomaly detection: identifies unusual API calls or planning step sequences [4], (ii) Goal consistency monitoring: ensures that the agent's high-level goals remain aligned with the intended parameters [19], (iii) Inter-agent behaviour correlation: detects abnormal influence propagation across multi-agent systems [14]

3) Scoped Autonomy and Sandboxing: The approach limits the blast radius of agent actions, thus preventing small misalignments from escalating into system-wide failures. Techniques include: (i) Using sandboxed execution environments for tool or API interactions [19], [23], (ii) Defining scoped permissions that dynamically adjust based on task requirements [23], (iii) Simulation-based testing, where potential plans are evaluated in a virtual environment before deploying them to production [19].

4) Human-in-the-Loop (HITL) Oversight: HITL is an important approach, especially for high-impact operations. The approach does not alter autonomous behaviour, but it in-

troduces selective oversight. It allows agents to work independently within safe boundaries, while escalating high-risk decisions to human operators [23], [40]. For example, if a user wishes to book a flight for less than £300, the system can search for available options. However, if the system were about to perform a high-risk task (such as paying for the booking), HITL oversight could be helpful, for instance, the system notices that the payment would be above £300. Thus, selective oversight can serve as an additional layer of protection, mitigating situations that may lead to adverse or unintended consequences.

5) Scalable Oversight: As mentioned above, using HITL oversight for selective operations can be a good defence. However, in agentic AI systems, this can also create challenges for humans [45], [46], as agents operate rapidly and dynamically spawn multiple sub-agents, making it cumbersome to monitor or approve each one individually. Hence, as a good defensive strategy, scalable oversight can be employed, which involves AI checking at AI. This method can protect systems, as one agent can judge over decisions taken by another agent at runtime (which can be missed by humans) [45], [46].

6) Memory Integrity and Verification: Some techniques can be used to prevent memory poisoning [41]: (i) Cryptographic primitives can be used to ensure the integrity of the stored memory, (ii) Memory snapshots with versions included can help the system to rollback to verified safe states.

7) Identity and Access Control Enhancements: To secure agents as digital identities, the following ways can be used [12], [20]: (i) Authenticate continuously using cryptographic primitives, (ii) Policy-based access tokens can be incorporated to limit privileges dynamically, (iii) Audit logging and non-repudiation mechanisms can be introduced to track agent actions across their lifecycle.

C. Challenges

While many security mitigation strategies have been proposed for agentic AI systems, their practical integration into existing standards and frameworks remains limited [11], [18], due to several challenges.

1) Lack of Standardisation and Formal Specifications: Existing frameworks, such as NIST AI RMF, TRiSM, and AGENTSAFE, were primarily developed keeping the notion of non-agentic AI systems as the center [11], [12], [18]. They majorly focused on: model evaluation metrics, data privacy, audit logging, and access control for human users. But agentic AI systems introduce a very different environment, which creates new attack surfaces not addressed by current standards [14]. For instance, (i) Runtime governance for multi-step executions has no formal definition in existing policy frameworks, (ii) Behavioural anomaly detection is context-dependent and cannot easily scale across different agent architectures, (iii) Autonomous agents require continuous monitoring and dynamic privilege management that further requires real-time orchestration mechanisms, which are absent in conventional AI governance standards. This absence of formalised specifications makes integrating mitigation strategies non-trivial since there are no universally accepted baselines or compliance metrics for autonomous behaviours [23].

2) Complexity of Multi-Agent and Dynamic Environments: Agentic AI systems frequently operate in multi-agent environments where agents coordinate across tools, APIs, and even physical actuators. Mitigation strategies at runtime must account for inter-agent influence propagation [14], cascading failures from compromised agents [14], [19], and dynamic modification of agent capabilities [23].

Existing standards assume isolated model behaviour rather than an interactive behaviour [12]. Integrating scoped autonomy, behavioural monitoring, and identity management in such dynamic ecosystems is challenging because (i) policies must adapt in real-time to evolving threats, (ii) risk assessment needs to model multi-agent interactions, which is computationally tricky, and (iii) current auditing frameworks do not support distributed accountability in agent networks [18], [20].

3) Trade-offs Between Autonomy and Security: The main challenge is how to balance the benefits of autonomy with security constraints. There are many mitigation strategies, such as HITL oversight or sandboxing, which introduce latency, reduce flexibility, and add operational complexity [40]. However, runtime enforcement may prevent agents from executing complex actions, limiting adaptability in novel situations [18]. Additionally, HITL oversight may slow down real-time incident response, reducing the operational advantage of agentic autonomy [40]. Finally, scoped permissions and sandboxing will require a fine-grained understanding of context-specific requirements, which is often unavailable or hard to generalise [23], [47]. Today's standards and frameworks do not consider how these trade-offs can be handled [42].

4) Integration Across Heterogeneous Systems: Agentic AI systems often interact with a wide range of environments and domains [19]. Each domain has its own security protocols and standards, which further create integration challenges, including: (i) behavioural anomaly detection models need cross-domain training to avoid false positives [47], (ii) memory verification mechanisms must maintain integrity while being compatible with diverse storage formats and tools [41], and (iii) runtime governance must be consistent with different authentication mechanisms, APIs, and network configurations as per requirements [12], [20].

Existing frameworks are not designed for cross-domain orchestration, making comprehensive integration of mitigation strategies an ongoing research challenge [7], [14].

5) Limitations in Verification and Assurance: Ensuring that mitigation strategies are effective, provable, and reliable is difficult due to the complex behaviour of autonomous agents that cannot be predicted, as well as attackers' evolving ways to bypass runtime constraints, memory checks, and behavioural monitoring [5], and a lack of formal verification tools for goal-directed, multi-step planning in agentic AI [47].

D. Current Research Directions

To address the challenges mentioned above, the following approaches are currently being explored.

1) Frameworks for Formal Runtime Governance: Researchers are developing policy-as-code systems to support real-time enforcement of constraints on agents' actions, which can help organisations to automate, manage, and enforce policies via Continuous Integration/Continuous deployment (CI/CD) pipelines [18]. To balance autonomous behaviour and safety, researchers are studying and trying to combine formal verification techniques with adaptive policies [47].

2) Risk Modelling for MAS: Researchers are investigating ways to cascade failures in MAS using simulation and then measuring the impact using network science and causal modelling techniques, such as the Load-Capacity Model, and Latent Graph Models, which may support inter-agent monitoring and anomaly detection [14].

3) Supervisory Control, Adaptive HITL and Scalable Oversight: Agentic AI systems and hybrid models use selective human oversight. They employ thresholds and risk-scoring methodologies to escalate only high-risk decisions to human oversight [40]. Current work explores approaches to automate coordination between agent autonomy and human control to optimise safety without impacting efficiency [40], [45], [46].

4) Secure Memory and Knowledge Management: Researchers are finding solutions using cryptographic memory commitments, tamper-evident logging, and verifiable learning updates for agentic AI [41], which will help ensure that agents are unable to drift in goals without detection, auditors verify that memory changes happened only after authorised updates, and verify that logs are not edited without breaking cryptographic integrity [41].

5) Identity and Access Control for Autonomous Agents: Current work focuses on developing mechanisms that dynamically manage credentials, provide continuous authentication, and implement blockchain-inspired auditing for agents [12], [20]. These mechanisms help prevent or limit threats such as persistent privilege abuse, prompt injection, and tool hijacking in autonomous systems. This research aims to extend traditional RBAC and Attribute-Based Access Control (ABAC) models to cover autonomous and evolving identities [20].

6) Simulation-Based Testing and Verification: Researchers are investigating how to simulate multi-agent scenarios with adversarial injections to stress-test mitigation strategies before deployment [19], [42], enabling frameworks to evolve iteratively and incorporate lessons learned from realistic threat models [19].

7) Trusted Delegation and OAuth Extensions for Agentic AI: Major research is underway aimed at enabling the safe and trustworthy delegation of authority in multi-agent collaboration and in accessing sensitive APIs [43]. For instance, traditional OAuth frameworks are being extended to agentic contexts to allow: (i) delegation of entitlements from humans to agents (user-to-agent) or between agents (agent-to-agent) or device to agents [43], [48], (ii) authority separation, i.e., authority to use delegated entitlements versus the authority to delegate further, thus, minimising risk of privilege escalation [43], [49], and (iii) context-aware access policies that ensures agents act only within approved operational scopes [48], [50]. Current research is about exploring formal semantics, cryptographically verifiable tokens, and runtime policy enforcement to provide trust guarantees for delegated actions in autonomous systems [43], [47], [49].

Integrating mitigation strategies for agentic AI is an ongoing, multifaceted research challenge. While progress is substantial, formal adoption into standards remains limited, highlighting the gap between cutting-edge research and operational frameworks.

VI. Discussion

Despite all of the potential security risks associated with autonomous behaviour, agentic AI still has capabilities that non-agentic systems cannot match [3]. For instance, in the cybersecurity domain, agentic AI systems can perform continuous monitoring, adaptive threat hunting, and real-time incident response at a broader scale, going beyond human capacity, such as identifying patterns and mitigating threats that would otherwise go undetected [8].

The use of agentic AI does not come without challenges. The main concern here is not whether these systems should be used, but rather how their autonomous behaviour can be constrained, monitored, and aligned with human intent [47]. Without proper restrictions and strategies, agentic systems could lead to various issues [12]. As discussed in Section V-B, with appropriate mitigation and security mechanisms, these systems can implement various tasks. Finally, the research and deployment of agentic AI are closely interconnected, as achieving a balance between capability and control is crucial [4]. By integrating robust security, governance, and accountability measures from the outset, it is possible to fully leverage the capabilities of agentic AI [19].

Limitations. This research focuses primarily on software-based agentic systems, for example, LLM-driven agents operating over APIs and enterprise workflows. Physical agentic systems, such as robotics and cyber-physical systems, introduce additional attack surfaces, such as sensor spoofing, actuator misuse, and real-world safety constraints that are outside the scope of this work.

VII. Conclusion

Agentic AI represents a transformation in the AI world, i.e., moving systems from passive tools to autonomous agents. This shift requires robust and adaptive security, governance, and ethical frameworks, which cannot be fully addressed using traditional AI security paradigms alone. By examining motivations, risks, and research gaps, this paper lays the groundwork for future efforts to develop secure and trustworthy agentic AI systems.

Future work based on this research should focus on formalising the threat model for agentic AI, including developing models that can capture the agent's interaction, memory persistence, planning depth, and tool execution. These models could help verify security properties such as bounded autonomy, safe delegation, and resistance to goal manipulation. The second objective is to implement and evaluate agentic authorisation

mechanisms, especially OAuth extensions and agent-specific delegation tokens. Future work should involve building prototype systems that would integrate the extensions into real agent frameworks, followed by adversarial testing to assess their effectiveness against various possible threats across agent hierarchies.

REFERENCES

[1] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges," *Information Fusion*, vol. 126, p. 103599, 2026.

[2] M. Wooldridge, *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[3] A. Bandi, B. Kongari, R. Naguru *et al.*, "The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges," *Fut. Internet*, vol. 17, no. 9, 2025.

[4] V. Bodepudi, N. Katnapally, V. Velaga *et al.*, "Agentic AI and reinforcement learning: towards more autonomous and adaptive AI systems," *J Educ Teach Train*, vol. 11, no. 1, pp. 177–193, 2020.

[5] H. Sarjoughian, B. Zeigler, and S. Hall, "A layered modeling and simulation architecture for agent-based system development," *Proceedings of the IEEE*, vol. 89, no. 2, pp. 201–213, 2001.

[6] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recog.*, vol. 84, p. 317–331, 2018.

[7] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *preprint arXiv:1412.6572*, 2014.

[8] M. Abou Ali, F. Dornaika, and J. Charafeddine, "Agentic AI: a comprehensive survey of architectures, applications, and future directions," *Artificial Intelligence Review*, vol. 59, no. 1, 2025.

[9] S. Yao, J. Zhao, D. Yu *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," *preprint arXiv:2210.03629*, 2023.

[10] K. Greshake, S. Abdelnabi, S. Mishra *et al.*, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," *preprint arXiv:2302.12173*, 2023.

[11] F. Perez and I. Ribeiro, "Ignore Previous Prompt: Attack Techniques For Language Models," *preprint arXiv:2211.09527*, 2022.

[12] (2023) NIST, AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[13] (2023) OpenAI.GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[14] A. Zou, Z. Wang, N. Carlini *et al.*, "Universal and Transferable Adversarial Attacks on Aligned Language Models," *preprint arXiv:2307.15043*, 2023.

[15] M. Naeem, S. T. H. Rizvi, and A. Coronato, "A Gentle Introduction to Reinforcement Learning and its Application in Different Fields," *IEEE Access*, vol. 8, pp. 209 320–209 344, 2020.

[16] G. Zhang, H. Geng, X. Yu *et al.*, "The Landscape of Agentic Reinforcement Learning for LLMs: A Survey," *arXiv:2509.02547*, 2025.

[17] V. S. Narajala and O. Narayan, "Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents," *preprint arXiv:2504.19956*, 2025.

[18] A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions," *Expert Systems with Applications*, vol. 240, p. 122442, 2024.

[19] A. K. Pati, "Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications," *IEEE Access*, vol. 13, pp. 151 824–151 837, 2025.

[20] N. Goel and N. Gupta, "Extending STRIDE and MITRE ATLAS for AI-Specific Threat Landscapes," *Well Testing*, vol. 34 No. S1 (2025), pp. 181–196, 2025.

[21] K. J. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for ai agents," *arXiv preprint arXiv:2506.12469*, 2025.

[22] J. Reavis. (2026) Leveling Up Autonomy in Agentic AI. https://cloudsecurityalliance.org/blog/2026/01/28/levels-of-autonomy.

[23] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," *IEEE Access*, vol. 13, pp. 18 912–18 936, 2025.

[24] Business Systems UK. Proactive AI vs Reactive AI: Understanding the Difference. https://bslgroup.com/proactive-ai-vs-reactive-ai-understanding-the-difference/.

[25] D. Dutta. (2025) From Chatbots to Autonomous Agents: Understanding Agentic AI and Its Security Risks. https://medium.com/@dipikanta.dutta/from-chatbots-to-autonomous-agents-understanding-agentic-ai-and-its-security-risks-b0aa95a7688a.

[26] C. Al-Dhubaib and I. Lee. (2025) From "Agents" to Autonomy: A Practical Framework for Agentic AI (Levels 1–5). https://datasaur.ai/blog-posts/from-agents-to-autonomy-a-practical-framework-for-agentic-ai-levels-1-5.

[27] D. Schofield. (2025) Defining the Autonomous Enterprise: Reasoning, Memory, and the Core Capabilities of Agentic AI. https://unstructured.io/blog/defining-the-autonomous-enterprise-reasoning-memory-and-the-core-capabilities-of-agentic-ai.

[28] M. Varavooru. (2026) Guardrails for Agentic AI — The Invisible Infrastructure That Determines Whether AI Scales or Fails. https://www.linkedin.com/pulse/guardrails-agentic-ai-invisible-infrastructure-scales-varavooru-stute.

[29] B. Blacet. (2025) What Does Agentic Mean? Understanding Agentic AI and Why It Matters for Enterprise Work. https://www.moveworks.com/us/en/resources/blog/what-does-agentic-mean.

[30] (2025) The Six Levels of Agentic Behavior. https://www.vellum.ai/blog/levels-of-agentic-behavior?utm_source=google&utm_medium=organic.

[31] J. Mackinlay and A. Veenendaal. (2025) AI Agents in KYC Compliance. https://www.blueprism.com/resources/blog/kyc-ai-agents-compliance/.

[32] (n.d.) Enterprise Agentic Automation: How AI Agents Transform Business Processes. https://www.informatica.com/resources/articles/enterprise-agentic-automation.html.

[33] S. Zhu, C. Fang, D. Larson *et al.*, "Compliance Brain Assistant: Conversational Agentic AI for Assisting Compliance Tasks in Enterprise Environments," *preprint arXiv:2507.17289*, 2025.

[34] A. Verhagen, A. Luget, O. Conjeaud *et al.* (2025) How agentic AI can change the way banks fight financial crime. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-agentic-ai-can-change-the-way-banks-fight-financial-crime.

[35] J. Kaur. (2025) Automating the KYC Process with Agentic AI. https://www.akira.ai/blog/automating-kyc-process-with-agentic-ai.

[36] G. Wang, Y. Xie, Y. Jiang *et al.*, "Voyager: An Open-Ended Embodied Agent with Large Language Models," *preprint arXiv:2305.16291*, 2023.

[37] MITRE. (2025) ATLAS. https://atlas.mitre.org/.

[38] A. Litan. (2024) Tackling Trust, Risk and Security in AI Models. https://www.gartner.com/en/articles/ai-trust-and-ai-risk.

[39] R. Khan, D. Joyce, and M. Habiba, "AGENTSAFE: A Unified Framework for Ethical Assurance and Governance in Agentic AI," *preprint arXiv:2512.03180*, 2025.

[40] A.-R. O. Ottun and H. Flores, "Trustworthy AI in Practice: A Comprehensive Review of Human Oversight and Human-in-the-Loop Approaches," *Authorea Preprints*, 2025.

[41] B. D. Sunil *et al.*, "Memory Poisoning Attack and Defense on Memory Based LLM-Agents," *preprint arXiv:2601.05504*, 2026.

[42] A. K. Pakina, "Alignment Drift as a Security Threat: Detecting and Mitigating Misaligned AI Behavior in Regulated Systems," *International Journal of Innovative Science and Research Technology*, p. 1856, 2025.

[43] A. Goswami, "Agentic JWT: A Secure Delegation Protocol for Autonomous AI Agents," *preprint arXiv:2509.13597*, 2025.

[44] P. McDaniel, N. Papernot, and Z. B. Celik, "Machine Learning in Adversarial Settings," *IEEE Security & Privacy*, vol. 14(3):68-72, 2016.

[45] R. Yin, T. Ishida, and M. Sugiyama, "Scalable Oversight via Partitioned Human Supervision," in *The Fourteenth International Conference on Learning Representations*, 2026.

[46] M. Bronikowski, "Scalable Oversight in Multi-Agent Systems: Provable Alignment via Delegated Debate and Hierarchical Verification," in *Open Conference of AI Agents for Science*, 2025.

[47] K. Huang, V. S. Narajala, J. Yeoh *et al.*, "A Novel Zero-Trust Identity Framework for Agentic AI: Decentralized Authentication and Fine-Grained Access Control," *preprint arXiv:2505.19301*, 2025.

[48] (2025) OAuth 2.0 Extension: On-Behalf-Of User Authorization for AI Agents. https://www.ietf.org/archive/id/draft-oauth-ai-agents-on-behalf-of-user-01.html.

[49] (2025) AAuth - Agentic Authorization OAuth 2.1 Extension. https://www.ietf.org/archive/id/draft-rosenberg-oauth-aauth-00.html.

[50] T. South, S. Marro, T. Hardjono *et al.*, "Authenticated Delegation and Authorized AI Agents," *preprint arXiv:2501.09674*, 2025.