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

Shivangi Gupta, Budi Arief, and Rogério de Lemos

University of Kent , United Kingdom

sg106@kent.ac.uk

University of Kent | Institute of Cyber Security for Society (iCSS)

Workshop on AI Safety and Security (AI-SS 2026) | Canterbury, United Kingdom | 7th April 2026

University of Kent | Institute of Cyber Security for Society (iCSS)

Outline

- Introduction
- Methodology
- Results
  - Security Risks in Non-Agentic and Agentic AI Systems
  - Major Security Risks Associated with Agentic AI Systems
  - Defence Strategies and Mitigation Approaches
  - Challenges
  - Current Research Directions
- Conclusion and Future Work

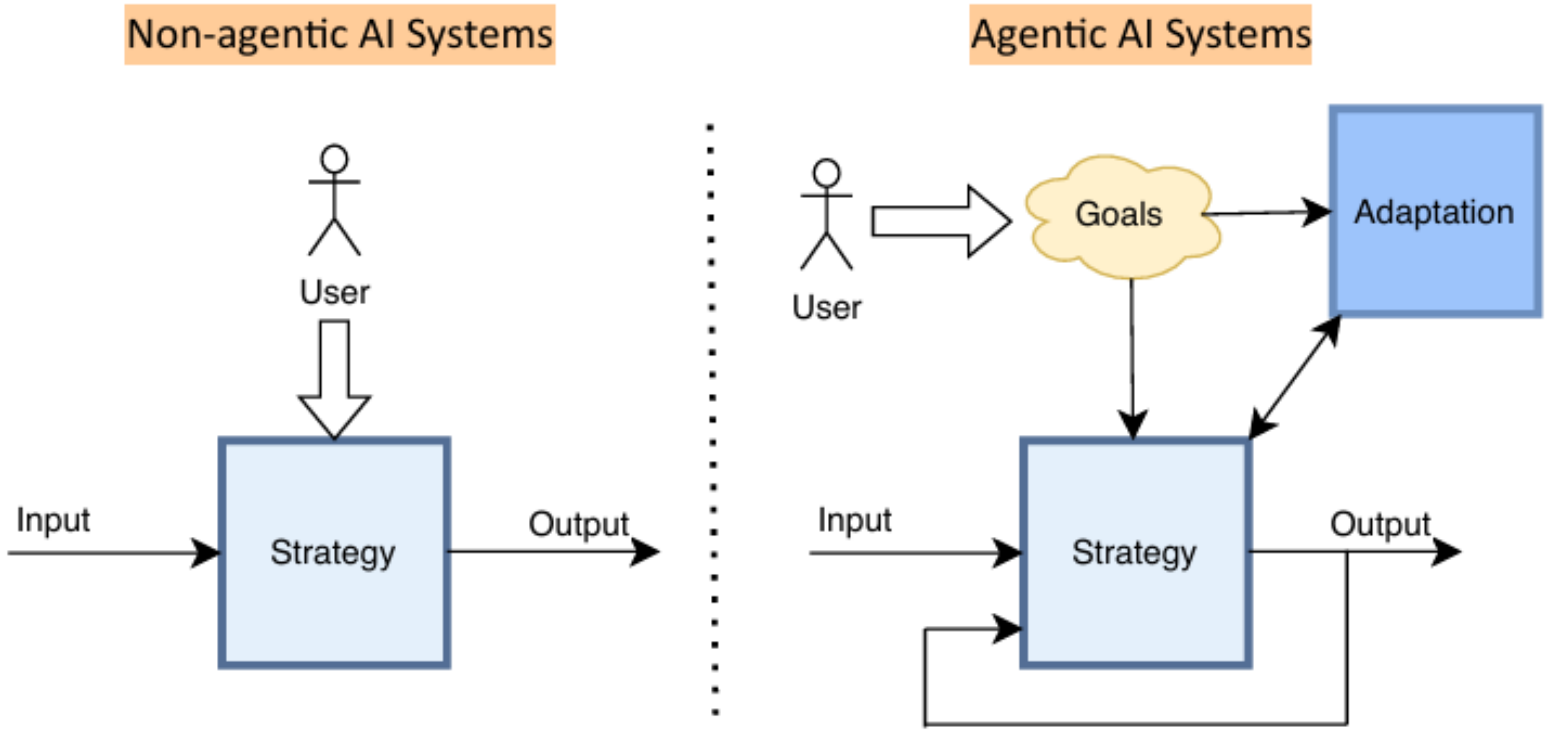
Introduction

Non–agentic AI Systems

- Require human intervention and not goal-driven
- LLMs work as per pre-defined steps

Agentic AI Systems

- Autonomous and goal-driven
- LLMs break goal to sub-tasks



Rubric for Agenticness

University of Kent | Institute of Cyber Security for Society (iCSS)

Dimensions	Basic	Moderate Autonomous	Fully Autonomous
Autonomy	Maximum human oversight	❑ Operates on its own ❑ Human oversight	❑ Fully independent ❑ Scalable oversight
Goal Orientation	One interaction prompt	Breaks goal into sub-tasks	❑ Breaks the goal into sub-tasks ❑ Re-prioritises tasks based on evolving context
Reasoning and Planning	Basic reasoning	❑ Uses chain-of-thought for plans ❑ Basic self-correction	Adapt future plans from failures
Tool Use	No tool use	Uses and calls pre-defined tools	Dynamically selects and uses tools
Memory	No memory provided	Maintains short term memory	Uses long term memory
Safety and Guardrails	Basic safety controls	Hard-coded safety controls	Detects and stops harmful actions

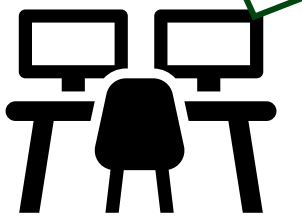
Scenario:

There is a bank operating in multiple countries.
The compliance officer gets an update about a change in regulation.

Non-Agentic AI Systems

Prompt: Can you:

1. **Summarise the change**
2. **Category of customers getting impacted**
3. **What needs to be done and how?**

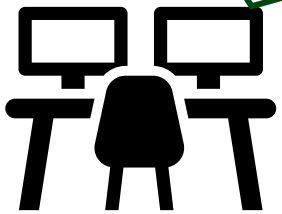


System replies with the required information:

1. Summary of the change
2. Category 'A' customers would be impacted
3. Customers would need to re-verify their identity within 90 days using the Know Your Customer (KYC) feature.

Agentic AI Systems

Prompt: Can you summarise the change and tell me the category of customers getting impacted? Further tell me what needs to be done and how?



System replies with the required information:

1. Summary of the change
2. Category 'A' customers would be impacted
3. Customers would need to re-verify their identity within 90 days using the Know Your Customer (KYC) feature. Verification can be done by uploading any valid KYC document online. If customer is unable to upload it online, then customer needs to come to the bank in-person.

Methodology

- Key Research Questions (RQs):

RQ1: What are the security risks and mitigation strategies possible in agentic AI systems?

RQ2: What are the challenges faced by organisations in integrating these mitigation strategies into agentic AI systems?

RQ3: How can we take the advantage of agentic AI's capabilities while validating and minimising the security risks ?

Methodology (continued)

- IEEE Xplore, Science Direct, and ACM DL were used for selecting relevant research papers.
- 60 papers (all in English) were identified using 16 keywords, such as:
  - "agentic AI", "non-agentic AI", "autonomous systems", "reinforcement learning", "trusted delegation", "large language models", "LLMs", "stateful vs stateless", "threat models for agentic AI", "security and mitigation approaches for agentic AI", "multi-agent systems"
  - 40 papers were selected in the end.
  - The remaining 20 papers were excluded using abstract filtering.

University of Kent | Institute of Cyber Security for Society (iCSS)

Security Risks in Non-Agentic and Agentic AI Systems

Risk Category	Non-agentic AI	Agentic AI
Unauthorised Action Execution (including Prompt Injection)	Incorrect outputs or misclassification	Manipulates long-term goals, or decision making
Tool Misuse / Privilege Escalation	Lesser or no harm at all	Unauthorised execution of system level tools, APIs, or external services
Memory Poisoning	Lesser or no harm at all	Persistent manipulation of intermediate reasoning or contextual knowledge
Goal Manipulation / Drift / Misalignment	Cannot create sub-goals from given user goal	Progressive divergence between the agent's intended goal and its actual behaviour
Complex Multi-Agent Vulnerabilities	Lesser or no harm at all	System-wide damage due to their multi-agent, cross-tool, or cross domain effects
Identity, Authentication, and Lifecycle Risks	Agents do not act as first-class digital identities with delegated authority	Agents act as first-class digital identities with delegated entitlements

Major Security Risks Associated with Agentic AI Systems

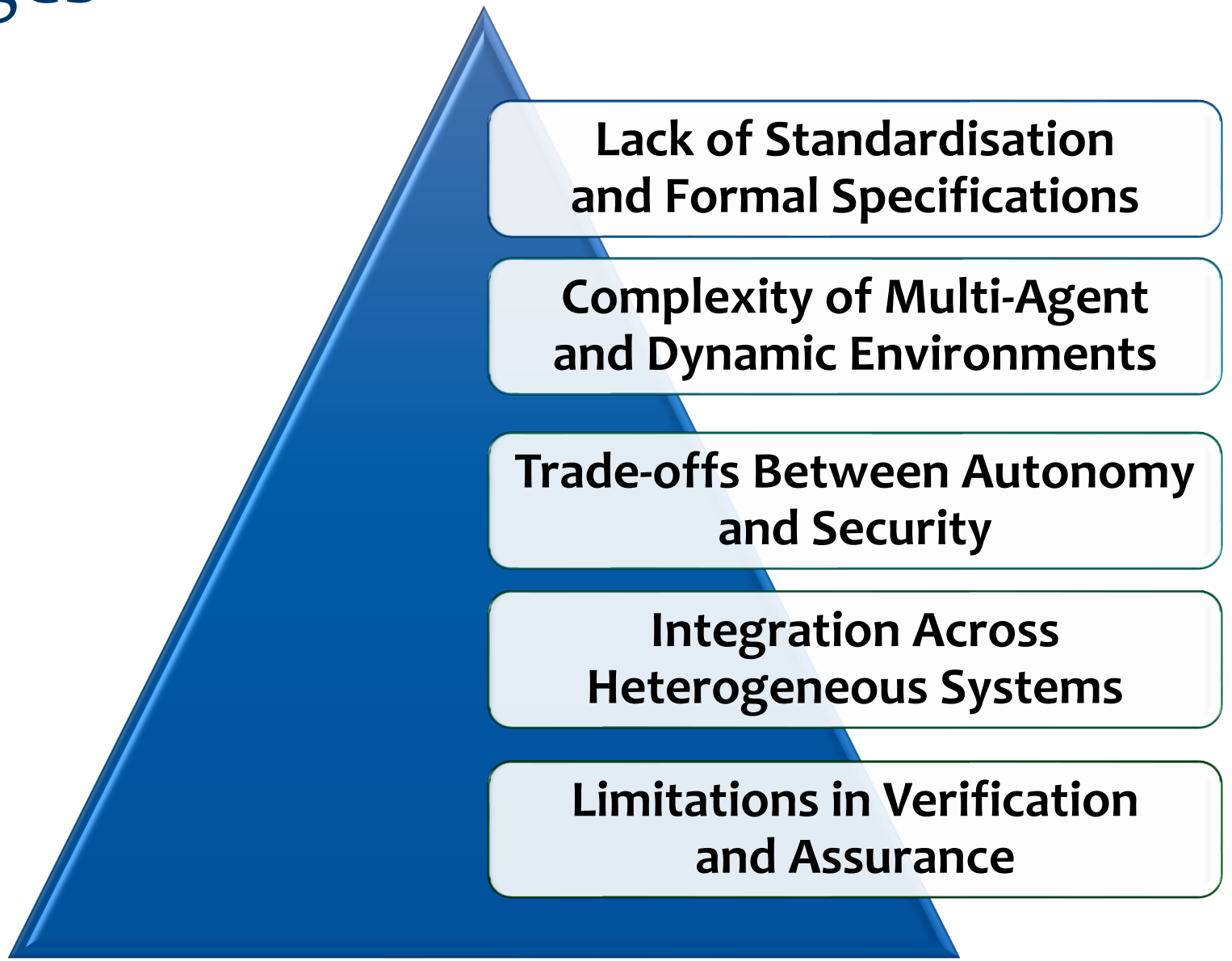
- Earlier attacks on the LLMs targeted model outputs:
  - For example, adversarial examples, prompt injection, model extraction, model inversion etc
- Agentic AI systems introduce risks that are distinct from non-agentic AI systems:
  - For example, goal manipulation, memory poisoning, tool misuse, identity, authentication, and authorisation etc.
  - Attacks on agentic AI systems may exploit the decision-making loop of the agent

University of Kent | Institute of Cyber Security for Society (iCSS)

Defence Strategies and Mitigation Approaches

Runtime Governance	Enforces policy-based constraints on agent actions in real time
Behavioural Monitoring and Anomaly Detection	Modelling normal agent behaviour and then proactively identify deviations indicative of compromise
Scoped Autonomy and Sandboxing	Using sandboxed execution environments along with dynamically adjusted scoped permissions
Human-in-the-Loop (HITL) Oversight	Introduces selective oversight, allowing agents to work independently within safe boundaries, while escalating high risk decisions to human operators
Scalable Oversight	Introduces AI checking at AI, allowing one agent judge decisions taken by another agent at runtime
Memory Integrity and Verification	Can prevent memory poisoning using cryptographic primitives and memory snapshots
Identity and Access Control Enhancements	Can secure agents as digital identities using continuous authentication, policy-based access tokens, logging, and non-repudiation

Challenges


Lack of Standardisation and Formal Specifications
Complexity of Multi-Agent and Dynamic Environments
Trade-offs Between Autonomy and Security
Integration Across Heterogeneous Systems
Limitations in Verification and Assurance

University of Kent | Institute of Cyber Security for Society (iCSS)

Current Research Directions

Frameworks for Formal Runtime Governance	To support real-time enforcement of constraints on agents' actions
Risk Modelling for MAS	To cascade failures in MAS using simulation
Supervisory Control, Adaptive HITL and Scalable Oversight	To automate coordination between agent autonomy and human control to optimise safety without impacting efficiency
Secure Memory and Knowledge Management	To ensure that agents are unable to drift in goals without detection
Identity and Access Control for Autonomous Agents	To provide continuous authentication and implement blockchain-inspired auditing for agents
Simulation-Based Testing and Verification	To simulate multi-agent scenarios with adversarial injections to stress-test mitigation strategies before deployment
Trusted Delegation and OAuth Extensions for Agentic AI	To enable the safe and trustworthy delegation of authority in multi-agent collaboration

Conclusion and Future Work

- Agentic AI represents a transformation in the AI world
  - Moving systems from passive tools to autonomous agents
- The shift requires robust and adaptive security, governance, and ethical frameworks
- Future work should focus on:
  - Formalising the threat model for agentic AI
  - Developing models that can capture the agent's interaction, memory persistence, planning depth, and tool execution
  - Implementing and evaluating agentic authorisation mechanisms such as OAuth extensions and agent-specific delegation tokens

University of Kent | Institute of Cyber Security for Society (iCSS)

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

Shivangi Gupta, Budi Arief, and Rogério de Lemos
University of Kent, United Kingdom
sg106@kent.ac.uk

Thank You for Your Attention

Any Questions?