LURKING IN THE SHADOWS

Sunil Agrawal *uncovers the hidden but critical security risk of unmonitored AI agents*

AI agents are moving rapidly from experimental tools to operational systems inside the enterprise. Unlike early generative AI applications, which primarily responded to user prompts, modern AI agents can initiate tasks, access entire corporate ecosystems and coordinate increasingly complex workflows. In many organisations, they're being used to retrieve internal knowledge, automate processes, generate reports and action multi-step processes across a plethora of software environments.

Their appeal is easy to understand. AI agents have the potential to free employees from repetitive, manual processes and drive significant productivity gains. It is no surprise, then, that adoption is accelerating across organisations. In fact, adoption is accelerating quickly. Research from McKinsey showed that by the end of 2025, 23 percent of organisations had already begun scaling an agentic AI system within at least one business function, while an additional 39 percent suggested they were experimenting with deployments. As we roll further into 2026, agentic systems feel less and less theoretical, and we will increasingly see them embedded in everyday enterprise operations. However, the speed of deployment is beginning to outpace the governance frameworks designed to manage them.

Security teams are now facing up to one of their greatest challenges to date: AI agents are being granted access to enterprise systems, sensitive data, and external interfaces faster than organisations can monitor or govern their behaviour. Without new forms of oversight, this shift risks introducing a category of security exposure that traditional controls were never designed to handle. It's not just that AI agents are powerful. It's that they operate entirely differently from conventional software. And those differences fundamentally reshape how enterprise risk must be managed.

Most organisations remain in the early stages of agent adoption, but usage is expanding across multiple functions. Recent research from Gravitee, the open-source leader in Agentic API and event management, found that across the US and UK alone, almost three-million AI agents have already been deployed, with nearly half of them (1.5-million) running without active oversight or security protocols. Gartner also found last year that 40 percent of enterprise applications will be integrated with task-specific AI agents by the end of 2026, while also suggesting over 40 percent of agentic AI projects will be cancelled by the end of 2027, due to (among other things) inadequate risk controls. The numbers reveal a clear trend. Enterprises are racing to deploy agentic capabilities, but security and risk control continue to derail deployments.

One related symptom of weak enterprise AI governance is the rapid growth of shadow AI. Employees are increasingly using publicly available AI tools to complete work tasks, often without the knowledge or approval of their organisation's IT or security teams. In many cases, these tools are connected to internal systems or used to process sensitive information. The result is the creation of parallel AI ecosystems operating outside established security frameworks. Within these environments, agents may access proprietary data, interact with internal applications, or generate outputs that influence decision-making — all without formal controls governing how those actions occur.

Employees are increasingly using publicly available AI tools without the knowledge or approval of their organisation's security teams

The consequences are already measurable. IBM's 2025 Cost of Data Breach Report found that: "97 percent of AI-related security breaches involved AI systems that lacked proper access controls. And most breached organisations reported they have no governance policies in place to manage AI or prevent shadow AI – the use of AI without employer approval or oversight". That does not mean shadow AI automatically leads to a breach. But it does increase organisational risk by reducing oversight and governance. And when breaches happen in those environments, the costs are often significantly higher – by an average of $670,000 compared with organisations with minimal shadow AI use.

From a security perspective, shadow AI represents more than a compliance issue. It reflects a broader shift in how technology is entering the enterprise: adoption is increasingly driven by user demand rather than centralised governance. AI agents only accelerate this trend by making automation easier to deploy and scale.

Shadow AI highlights what happens when AI adoption outpaces institutional control. But it is only one part of the broader challenge. Even when agents are officially approved and deployed inside enterprise workflows, organisations still need a way to govern how those systems behave once they are connected to sensitive data, enterprise applications, and external interfaces.

To understand why governing approved, operational AI agents requires a different approach, it's important to distinguish how they differ from conventional applications. Most enterprise software follows deterministic logic – a developer defines a set of rules, and the system executes tasks according to predefined instructions. AI agents behave differently. Instead of following fixed workflows, agents interpret objectives and dynamically determine the steps required to complete them. This allows them to perform sophisticated tasks, such as synthesising information across multiple systems,but this also introduces a vastly expanded attack surface.

AI AGENTS ARE BEING GRANTED ACCESS TO SENSITIVE DATA AND EXTERNAL INTERFACES

An AI agent may interact with dozens of APIs, databases and services in ways that developers did not explicitly script. Each connection expands the potential pathways through which data can flow and therefore the potential points of compromise. Compounding this challenge is a structural limitation within large language models themselves: they do not reliably separate instructions from data, meaning malicious content embedded in inputs can influence model behaviour. Security researchers often describe this scenario using what's known as the 'legal trifecta' of agent vulnerabilities, which comprises of access to sensitive data, exposure to untrusted external inputs and the ability to communicate or act externally. When an agent possesses all three of these capabilities simultaneously, it becomes an attractive target for prompt injection attacks, data exfiltration attempts, or manipulation through adversarial inputs.

But it's not just external threats that warrant attention. Many AI-related vulnerabilities originate from inside the organisation and, once again, AI agents amplify the risk associated with some of the behaviours we already see with generative AI use. Employees may upload sensitive documents, connect systems in insecure ways or create workflows that expose confidential information, but the risk amplification comes from agents that can act autonomously based on flawed or incomplete context.

In effect, an incorrectly configured agent can unintentionally perform actions that resemble insider threats: retrieving data, sharing information externally or executing commands that exceed its intended scope. This doesn't require malicious intent; it only requires a system that lacks sufficient guardrails.

Most enterprise security frameworks were designed around a simple assumption: humans initiate actions and systems enforce rules around those actions. Identity management verifies users. Access controls determine which resources they can reach. Network monitoring detects suspicious activity. But AI agents disrupt this model. Instead of waiting for human commands, agents

Picture credit: AdobeStock

may initiate tasks independently, coordinate with other systems and continuously update their behaviour based on new information.

While traditional logging systems can record what actions have occurred, they rarely capture the context behind those decisions, ie: why the agent took a particular action or whether it was influenced by manipulated input. Similarly, static policy rules often struggle to govern systems that dynamically reinterpret instructions. This means an agent might receive a prompt that appears legitimate, but subtly alters its behaviour in ways that conflict with established policies.

MANY AI-RELATED VULNERABILITIES ORIGINATE FROM INSIDE THE ORGANISATION

In these scenarios, verifying credentials or enforcing network boundaries simply isn't enough. Security teams must also evaluate behaviour, intent and context. In other words, governance must evolve from controlling access to monitoring decision-making.

Addressing these challenges requires a governance framework capable of observing, guiding and constraining approved AI agent behaviour across the enterprise. Without this evolution, AI agents will only become more autonomous as human managers lose visibility into what the technology is doing and why.

One useful lens for approaching the governance of operational AI agents is the AWARE model, which evaluates activity across several dimensions that together provide visibility into how autonomous systems operate. Rather than focusing on system access or network activity, the model encourages organisations to assess the behavioural environment in which agents function.

GOVERNING AUTONOMY

Much of the conversation around AI security still focuses on protecting the models themselves. In practice, the harder problem is governing what those systems do once they are embedded inside the enterprise. As these systems move from answering questions to executing tasks across the enterprise, they become part of the operational fabric of the business. They retrieve information, trigger workflows and interact with systems in ways that are often invisible to the teams responsible for securing them.

That is the real governance challenge. It's already quite clear that organisations will continue to roll out AI agents in full force, but security teams have to maintain a clear line of sight into how those systems behave once they are embedded in everyday operations. Autonomy may be powerful, but in security, autonomy without visibility is simply risk. The organisations that succeed will be the ones that build governance in from the very beginning, so innovation can scale safely and responsibly ●

Sunil Agrawal, Glean's Chief Information Security Officer, has over 20 years of experience in the field of security, with specific expertise in conceptualizing and delivering innovative and high-quality security solutions.

THE AWARE FRAMEWORK

Actor Intent: The first dimension focuses on who or what is acting and why. This involves evaluating whether an agent's actions align with legitimate business objectives and authorised workflows. By understanding the intended purpose of the agent and the scope of its permissions, organisations can distinguish productive automation from anomalous behaviour.

Work Context: The second dimension considers the sensitivity of the information involved and the nature of the task being performed. Reading data carries different risks than modifying or transmitting it externally. Governance frameworks must therefore account for the context in which actions occur, ensuring that security decisions reflect the criticality of the underlying data.

Autonomous Guardrails: Agents should also operate within clearly defined operational boundaries that prevent them from executing tasks outside their intended scope. Guardrails can include restrictions on external communications, limits on data access and policies that prevent high-risk actions from occurring without human validation.

Real-Time Risk Scoring: Given the dynamic nature of AI behaviour, risk evaluation must also occur in real time. Systems should be capable of identifying high-risk actions before they are executed, allowing organisations to intervene when behaviour deviates from expected patterns.

Ecosystem Observability: Finally, organisations must maintain visibility across the broader ecosystem in which agents operate. This includes tracing interactions across systems, reconstructing decision pathways and ensuring that agent actions remain auditable. Without this observability, security teams lose the ability to understand how automated decisions influence enterprise operations.

The speed of deployment of AI agents is beginning to outpace the governance frameworks designed to manage them