

## Hey! What's New? 2025-113

### Deploying Agentic AI with Safety and Security: A Playbook for Technology Leaders

According to a new report from McKinsey & Company, “business leaders are rushing to embrace agentic AI, and it’s easy to understand why. Autonomous and goal driven, agentic AI systems are able to reason, plan, act and adapt without human oversight – powerful new capabilities that could help organizations capture the potential unleashed by gen AI by radically reinventing the way they operate. A growing number of organizations are now exploring or deploying agentic AI systems, which are projected to help unlock \$2.6 trillion to \$4.4 trillion annually in value across more than 60 gen AI use cases, including customer service, software development, supply chain optimization and compliance. And the journey to deploying agentic AI is only beginning: just 1 percent of surveyed organizations believe that their AI adoption has reached maturity.”

But while agentic AI has the potential to deliver immense value, the technology also presents an array of new risks, says the report, “introducing vulnerabilities that could disrupt operations, compromise sensitive data, or erode customer trust. Not only do AI agents provide new external entry points for would-be attackers, but because they are able to make decisions without human oversight, they also introduce novel internal risks.”

It is up to technology leaders, including chief information officers (CIOs), chief risk officers (CROs), chief information security officers (CISOs), and data protection officers (DPOs), “to develop a thorough understanding of the emerging risks associated with AI agents and agentic workforces and to proactively ensure secure and compliant adoption of the technology. The future of AI at work isn’t just faster or smarter. It’s more autonomous. Agents will increasingly initiate actions, collaborate across silos and make decisions that affect business outcomes. That’s an exciting development – provided those agents are working with not just a company’s access but also its intent. In an agentic world, trust is not a feature. It must be the foundation.”

By operating autonomously and automating tasks traditionally performed by human employees, agentic AI adds an additional dimension to the risk landscape, the report notes. “The key shift is a move from systems that enable interactions to systems that drive transactions that directly affect business processes and outcomes. This shift intensifies the challenges around core security principles of confidentiality, integrity and availability in the agentic context, due to the additional potential of amplifying foundational risks, such as data privacy, denial of services, and system integrity.

The following new risk drivers transcend the traditional risk taxonomy associated with AI:

- Cross-agent task escalation. “Malicious agents exploit trust mechanisms to gain unauthorized privileges.”
- Synthetic-identity risk. “Adversaries forge or impersonate agent identities to bypass trust mechanisms.”
- Untraceable data leakage. “Autonomous agents exchanging data without oversight obscure leaks and evade audits.”

- Data corruption propagation. “Low-quality data silently affects decisions across agents.”

The report points out that “such errors threaten to erode faith in the business processes and decisions that agentic systems are designed to automate, undermining whatever efficiency gains they deliver. Fortunately, this is not inevitable. Agentic AI can deliver on its potential, but only if the principles of safety and security outlined are woven into deployments from the outset.”

To adopt agentic AI securely, it advises, “organizations can take a structured, layered approach. Below, we provide a practical road map that outlines the key questions technology leaders should ask to assess readiness, mitigate risks and promote confident adoption of agentic systems. The journey begins with updating risks and governance frameworks, moves to establish mechanisms for oversight and awareness, and concludes with implementing security controls.”

For the details, read the report at [Agentic AI security: Risks & governance for enterprises | McKinsey](#),