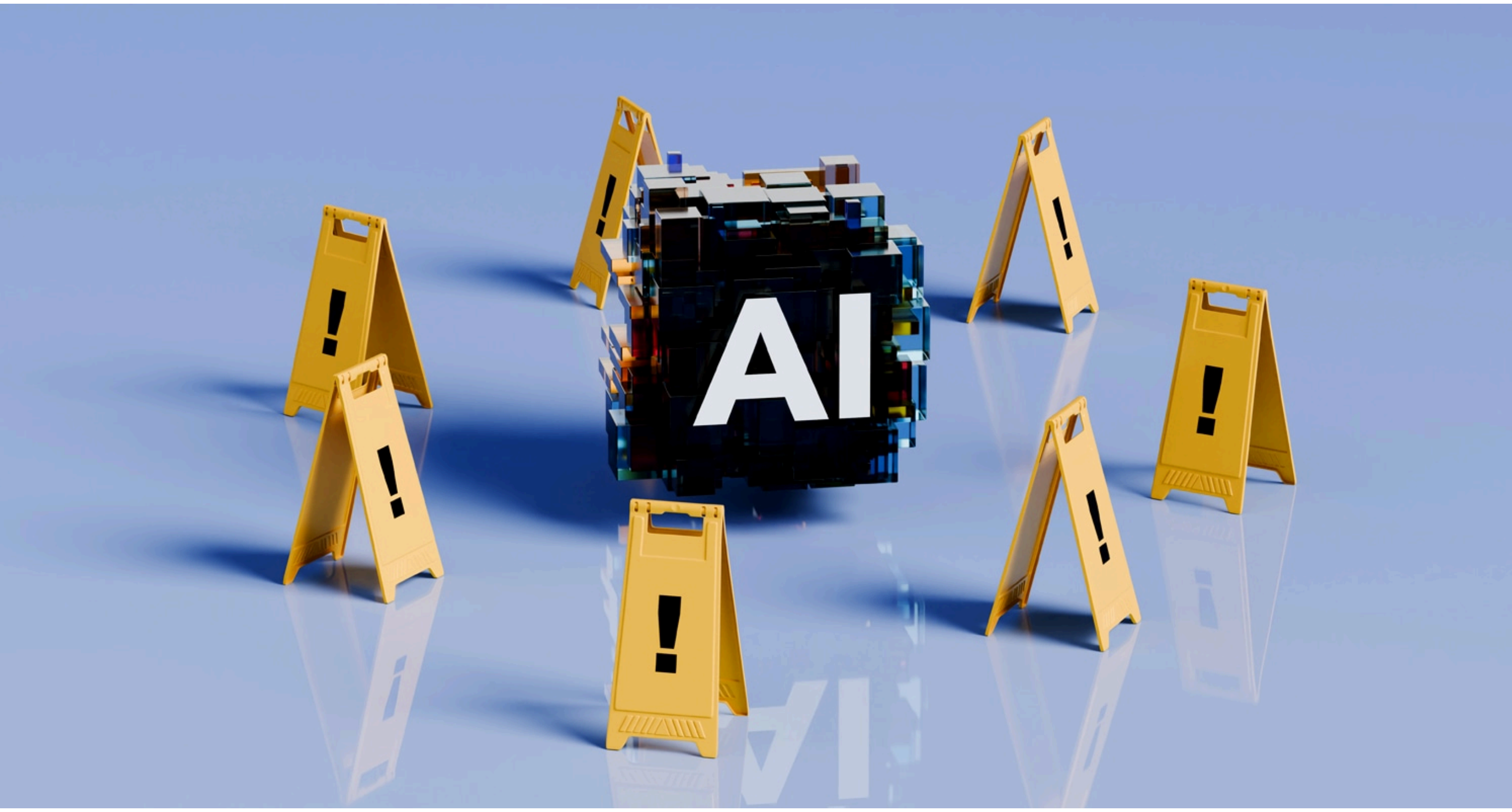
McKinsey Quarterly

Risk & Resilience Practice

Deploying agentic AI with safety and security: A playbook for technology leaders

Autonomous AI agents present a new world of opportunity—and an array of novel and complex risks and vulnerabilities that require attention and action now.

This article is a collaborative effort by Benjamin Klein, Charlie Lewis, and Rich Isenberg, with Dante Gabrielli, Helen Möllering, Raphael Engler, and Vincent Yuan, representing views from McKinsey's Risk & Resilience Practice.



October 2025

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.[1] And the journey to deploying agentic AI is only beginning: just 1 percent of surveyed organizations believe that their AI adoption has reached maturity.[2]

But while agentic AI has the potential to deliver immense value, the technology also presents an array of new risks—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. In cybersecurity terms, you might think of AI agents as "digital insiders"—entities that operate within systems with varying levels of privilege and authority. Just like their human counterparts, these digital insiders can cause harm unintentionally, through poor alignment, or deliberately if they become compromised. Already, 80 percent of organizations say they have encountered risky behaviors from AI agents, including improper data exposure and access to systems without authorization.[3]

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. (A review of early agentic AI deployments highlights six key lessons—from reimagining workflows to embedding observability—that can help organizations avoid some common pitfalls as they scale the new technology.[4]) 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.

Emerging risks in the agentic era

By operating autonomously and automating tasks traditionally performed by human employees, agentic AI adds an additional dimension to the risk landscape. 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[5]:

[1] "The promise and the reality of gen AI agents in the enterprise," McKinsey, May 17, 2024.
[2] Hannah Mayer, Lareina Yee, Michael Chui, and Roger Roberts, "Superagency in the workplace: Empowering people to unlock AI's full potential," McKinsey, January 28, 2025.
[3] *AI agents: The new attack surface; A global survey of security, IT professionals and executives*, SailPoint Technologies, May 28, 2025.
[4] Lareina Yee, Michael Chui, Roger Roberts, and Stephen Xu, "One year of agentic AI: Six lessons from the people doing the work," McKinsey, September 12, 2025.
[5] "Implementing generative AI with speed and safety," *McKinsey Quarterly*, March 13, 2024.

— *Chained vulnerabilities*. A flaw in one agent cascades across tasks to other agents, amplifying the risks.

Example: Due to a logic error, a credit data processing agent misclassifies short-term debt as income, inflating the applicant's financial profile. This incorrect output flows downstream to the credit scoring and loan approval agents, leading to an unjustified high score and risky loan approval.

— *Cross-agent task escalation.* Malicious agents exploit trust mechanisms to gain unauthorized privileges.

Example: A compromised scheduling agent in a healthcare system requests patient records from a clinical-data agent, falsely escalating the task as coming from a licensed physician. The agent then releases sensitive health data, resulting in unauthorized access and potential data leakage without triggering security alerts.

— *Synthetic-identity risk.* Adversaries forge or impersonate agent identities to bypass trust mechanisms.

Example: An attacker forges the digital identity of a claims processing agent and submits a synthetic request to access insurance claim histories. Trusting the spoofed agent's credentials, the system grants access, exposing sensitive policyholder data without detecting impersonation.

— *Untraceable data leakage.* Autonomous agents exchanging data without oversight obscure leaks and evade audits.

Example: An autonomous customer support agent shares transaction history with an external fraud detection agent to resolve a query but also includes unneeded personally identifiable information about the customer. Since the data exchange isn't logged or audited, the leakage of sensitive banking data goes unnoticed.

— *Data corruption propagation.* Low-quality data silently affects decisions across agents.

Example: In the pharmaceutical industry, a data labeling agent incorrectly tags a batch of clinical-trial results. This flawed data is then used by efficacy analysis and regulatory reporting agents, leading to distorted trial outcomes and potentially unsafe drug approval decisions.

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 below are woven into deployments from the outset.

Guiding principles for agentic AI security

To adopt agentic AI securely, 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.

Prior to agentic deployment

Before an organization begins using autonomous agents, it should ensure that it has the necessary safeguards, risk management practices, and governance in place for a secure, responsible, and effective adoption of the technology. Here are some key questions to consider:

— *Does our AI policy framework address agentic systems and their unique risks?* Answering this question starts with upgrading existing AI policies, standards, and processes—such as identity and access management (IAM) and third-party risk management (TPRM)—to cover the new capabilities of agentic systems. For instance, in the context of IAM, organizations should define roles and approval processes for agents to protect interactions with data, systems, and human users. Similarly, they should define and review the interactions of agentic solutions acquired from third parties with internal resources.

  Organizations must also grapple with the ever-changing nature of AI regulations. They can start by identifying the rules they are subject to. Article 22 of the European Union's General Data Protection Regulation (GDPR), for example, restricts the usage of AI by granting individuals the right to deny decisions based solely on automated processing. In the United States, sector-specific laws such as the Equal Credit Opportunity Act (ECOA) impose requirements on AI systems to prevent discrimination. Additionally, state-level initiatives, such as New York City's Local Law 144, mandate bias audits for automated employment decision tools, signaling a growing trend toward AI accountability. New AI-specific regulations, like the EU AI Act, are being adopted and will take full effect in the next three years. In this rapidly evolving regulatory landscape, in which many requirements remain unclear, a conservative approach—anticipating likely standards, such as human oversight, data protection, and fairness—can help organizations stay ahead and avoid costly compliance overhauls in the future.

For each agentic use case in an organization's AI portfolio, tech leaders should identify and assess the corresponding organizational risks, and, if needed, update their risk assessment methodology.

- *Is our risk management program equipped to handle agentic AI risks?* Enterprise cybersecurity frameworks—such as ISO 27001, the National Institute of Standards and Technology Cybersecurity Framework (NIST CSF), and SOC 2—focus on systems, processes, and people. They do not yet fully account for autonomous agents that can act with discretion and adaptability. To bridge this gap, organizations can revise their risk taxonomy to explicitly account for the novel risks introduced by agentic AI (exhibit).

  For each agentic use case in an organization's AI portfolio, tech leaders should identify and assess the corresponding organizational risks, and, if needed, update their risk assessment methodology to be capable of measuring risks within agentic AI. Without this transparency, risks arising from agentic AI threaten to become a black box even more than what we've seen with analytical or gen AI.

- *Do we have robust governance for managing AI across its full life cycle?* Establishing governance requires defining standardized oversight processes, including ownership and responsibilities within AI onboarding, deployment, and offboarding procedures; monitoring and anomaly detection tied to KPIs; defining triggers for escalations; and developing standards of accountability for agent actions. For each agentic AI solution in the portfolio, organizations should start by listing technical details—such as foundational model, hosting location, and data sources accessed—as well as the criticality of the use case, contextual data sensitivity, access rights, and interagent dependencies. Next, they should establish clear ownership of each use case, with human-in-the-loop oversight and responsible stakeholders for decision-making, security, and compliance, while also identifying and allocating capabilities to manage the risks.

Exhibit

The introduction of agentic AI requires organizations to update their risk taxonomies.

AI risks by enterprise risk category, illustrative (not exhaustive)

Financial	Operational	People	Regulatory	Reputational	Strategic
● AI cost overrun ● Algorithmic financial exposure ● Synthetic fraud and transaction risk	● Data corruption/ model poisoning ●● System drift/ misbehavior ● Systemic dependency/ lack of fallback	● Accountability ambiguity/loss of human oversight ● Deskilling ● Skill gaps ● Stress and resistance ● Workforce displacement	● Bias/ discrimination ● Lack of transparency/ explainability ● Noncompliance ● Unauthorized data use or disclosure	● Controversial or misled AI decisions ● Stakeholder distrust ● Perceived ethical violations	● Opaque decision influence ● Overreliance ● Strategic misalignment

Acceleration because of agentic AI

- ● **Chained vulnerabilities:** Strategies built on fragile multiagent architectures
- ● **Cross-agent task escalation:** Agents expand decision scope or delegate tasks beyond intent
- ● **Data corruption propagation:** Impact of low data quality is amplified by decision chains across agents
- ● **Synthetic identity risk:** Use of agents to simulate identities, generate fraud, or manipulate transactions
- ● **Untraceable data leakage:** Exchange of data between agents without oversight obscures data leaks

● *Gen AI risks not linked to novel agentic AI risk types*

McKinsey & Company

Prior to launching an agentic use case

Once the above foundational questions have been addressed and an agentic AI risk framework and policies are in place, organizations should develop a clear understanding of precisely what they are building, accounting for associated risks and compliance considerations for each project. Addressing the following questions can help ensure that their ambitions are matched by readiness:

Especially in the experimental or piloting stage, AI projects have a way of proliferating rapidly without adequate oversight, which can make it challenging to manage risks or enforce governance.

- *How can we maintain control of agentic initiatives and ensure that we have oversight over all projects?* Especially in the experimental or piloting stage, AI projects have a way of proliferating rapidly without adequate oversight, which can make it challenging to manage risks or enforce governance. Organizations should establish a clear, centrally steered, and business-aligned AI portfolio management system that ensures oversight by IT risk, information security, and IT compliance functions. This system should provide full transparency around business, IT, and security ownership; detailed descriptions of use cases; a list of the data provided to the agent for training, interaction (for example, connected APIs), or both; and the status of the data. The repository should also include all agentic systems that are currently in development, being piloted, or being planned by business units. This can help organizations avoid experimental and uncontrolled deployment of models with potentially unintended critical exposure points.

- *Do we have the capabilities to support and secure our agentic AI systems?* To help ensure the success of agentic AI pilots, organizations should assess their current level of skills, knowledge, and resources in relation to the agentic road map—including AI security engineering, security testing, threat modeling, and the skills required for governance, compliance, and risk management. They should then identify the skill and resource gaps that exist between agentic ambitions and security capabilities and launch awareness and educational campaigns to narrow such gaps—while defining critical roles based on the AI life cycle. For example, organizations lacking knowledge regarding AI threats will need to upskill security engineers on threat modeling of AI models and agents.

During the deployment of an agentic AI use case

Once use cases and pilots are up and running, organizations will need to ensure that the pilots are enforced by technical and procedural controls. These controls should be regularly reassessed to ensure that they remain relevant and effective as agentic systems are refined and scaled. Here are some key questions to consider :

- *Are we prepared for agent-to-agent interactions, and are those connections secure?* AI agents interact with not only human users but also other AI agents. It is essential that organizations secure these agent-to-agent collaborations, especially as multiagent ecosystems grow. Protocols to manage agentic interactions, such as Anthropic's Model Context Protocol, Cisco's Agent Connect Protocol, Google's Agent2Agent protocol, and IBM's Agent Communication Protocol, are under development but not yet fully mature. As tech leaders monitor protocol evolution, they should also ensure that interagent communications are authenticated, logged, and properly permissioned. Rather than wait for perfect standards, it's best to implement safeguards now and plan for upgrades as more secure protocols emerge.

- *Do we have control over who can use agentic systems and whether they are using them appropriately?* Access to models and resources needs to be monitored and secured. Identity and access management systems should apply not only to human users, but also to AI agents that interact with other agents, humans, data, and system resources. Organizations should define which users, human or AI, are authorized to access or interface with such resources and assets and under what conditions. They should also augment IAM with input/output guardrails to prevent agents from being misused, manipulated, or triggered into unsafe behavior through adversarial prompts or misaligned objectives. Additionally, organizations need to carefully manage the ways in which third-party agentic AI agents interact with internal resources to help ensure that they meet the same security, governance, and ethical requirements as internal systems.

- *Can we trace agents' actions and understand and account for their behavior?* Agentic systems should be created with traceability mechanisms in place from the outset.[6] That means recording not only the agents' actions but also the prompts, decisions, internal state changes, intermediate reasoning, and outputs that led to these behaviors. Such systems are essential for auditability, root cause analysis, regulatory compliance, and postincident reviews. Organizations should establish regular performance reviews to evaluate whether agents remain aligned with their intended purpose.

- *Do we have a contingency plan if an agent fails or behaves unexpectedly?* Even well-designed agents can fail, become corrupt, or be exploited. Before deployment, organizations should develop a contingency plan, with proper security measures in place, for every critical agent. That starts with simulating worst-case scenarios, such as agents that become unresponsive, deviate from the expected objective, are intentionally malicious, or escalate tasks without

[6] Carlo Giovine, Roger Roberts, Mara Pometti, and Medha Bankhwal, "Building AI trust: The key role of explainability," McKinsey, November 26, 2024.

authorization. Next, organizations should ensure that termination mechanisms and fallback solutions are available. Lastly, they should deploy agents in self-contained environments with clearly defined network and data access. This also allows for immediate isolation if needed.

By identifying and implementing effective controls, organizations can proactively mitigate agentic AI risks rather than reactively responding to them. For instance, maintaining a consistent AI agent portfolio alongside robust AI logging enables the monitoring of data exchanges between agents, thereby mitigating the risk of untraceable data leakage. Additionally, deploying an AI contingency plan and sandbox environment, in conjunction with IAM and guardrails, can effectively isolate an AI agent that attempts unauthorized privilege escalation through cross-agent task escalation.

Agentic security cannot be an afterthought

The agentic workforce is inevitable. As more companies adopt AI agents, new challenges for maintaining the confidentiality and integrity of data and systems will arise. Currently, decision-makers face a pivotal moment to balance business enablement with a structured approach to risk management for agentic security; after all, no one wants to become the first agentic AI security disaster case study. CIOs, CROs, and CISOs should promptly engage in essential discussions with their business counterparts to gain transparency about the current state of agentic AI adoption in the organization and start building the essential guardrails. Acting thoroughly and with intention now will help ensure successful scaling in the future.

Currently, agentic transactions remain digital, but the trajectory points toward an even more radical future, including embodied agents operating in the physical world. The implications for safety and security will become even more profound, making it all the more important to prepare a strong foundation today.

Benjamin Klein is a partner in McKinsey's Berlin office, **Charlie Lewis** is a partner in the Connecticut office, **Rich Isenberg** is a partner in the Atlanta office, **Dante Gabrielli** is a director of product management in the Philadelphia office, **Helen Möllering** is a consultant in the Munich office, and **Raphael Engler** is an associate partner in the Zurich office, where **Vincent Yuan** is a consultant.

This article was edited by Larry Kanter, a senior editor in the New York office.