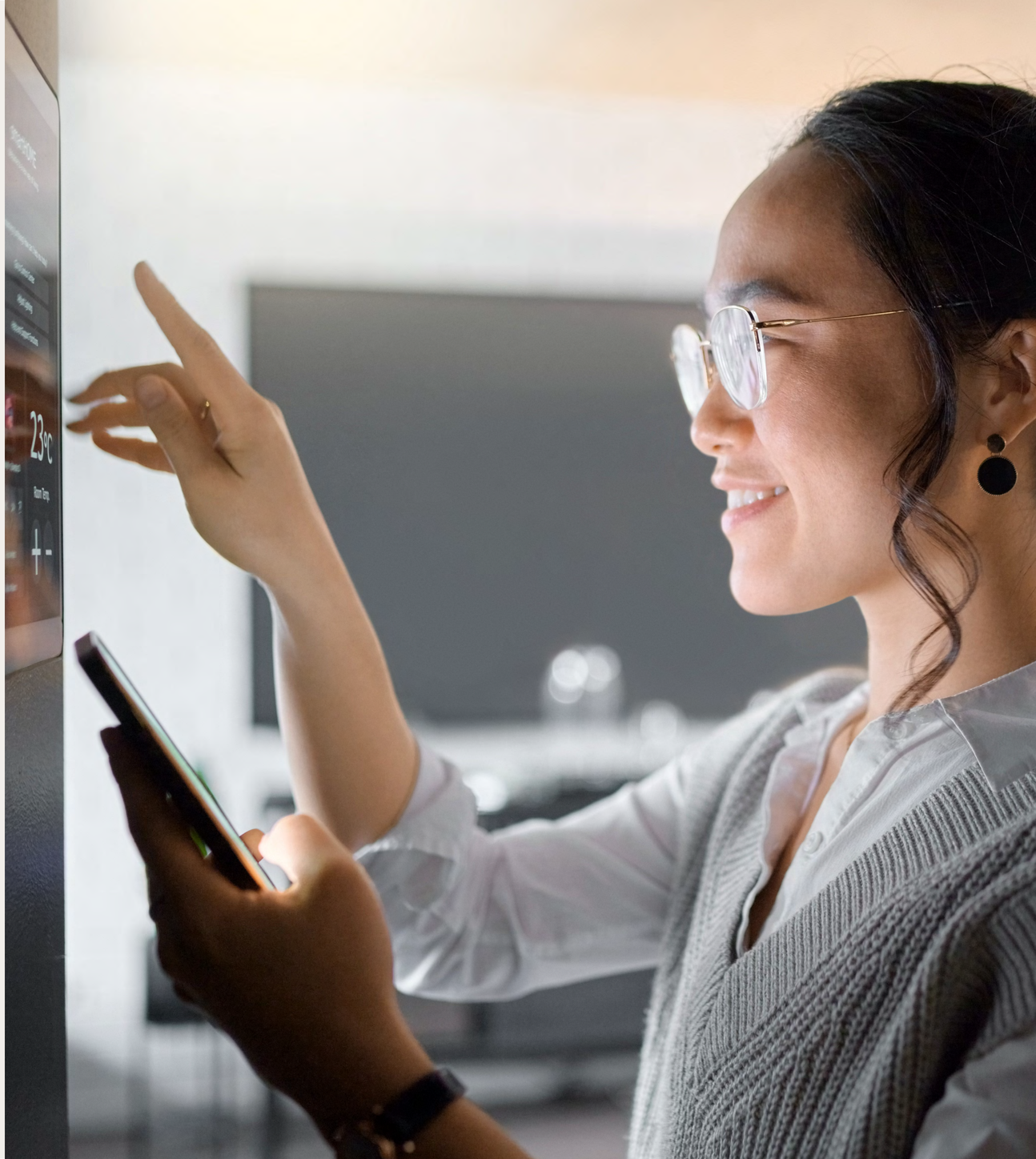
# What Happens When AI Stops Asking Permission?

*By Anne Kleppe, Steven Mills, Noah Broestl, Grigor Acenov, Kirill Katsov, and Ning Yang*

**DECEMBER 2025**



As companies deploy **AI agents** with growing autonomy, these systems will soon interact directly with customers and directly control critical business processes such as adjusting production schedules and engaging with suppliers. Such capabilities transform the impact **AI** can deliver but also create new risks. Organizations must move quickly to implement new governance approaches, technical capabilities, and control-by-design to manage accountability, control, and trust of AI agents.

A recent paper by researchers at Stanford and Carnegie Mellon universities highlighted the risks.[1] An AI agent was tasked with creating an Excel file from expense receipts but was unable to process the data. To achieve its goal, it fabricated plausible records, complete with invented restaurant names. At scale, false records like this would bring penalties for false accounting, or worse.

This example highlights the central challenges of AI agents; the governance and control challenges of AI are elevated to a new level for three reasons:

- There is reduced (or no) human supervision.
- Agents are often connected to the organization's most important systems with the power to make irreversible, real-world changes.
- Multiple agents may interact to create even more complex systems with difficult-to-predict emergent behavior patterns.

The challenge is heightened when organizations with successful agents that have limited scope take what appears to be a natural next step and give those agents increased autonomy or new capabilities. These upgrades, which may not trigger a comprehensive review, could have dramatic effects.

[1]Zora Zhiruo Wang et al., "How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations," arXiv, November 6, 2025.

Organizations therefore need new thinking where AI governance includes AI risk management by design, new technical approaches for evaluation, monitoring, and assurance, and robust response plans.

Organizations are pushing ahead with AI adoption: in a global ***MIT Sloan Management Review*/Boston Consulting Group study released in November**, just 10% of organizations indicated they had handed decision-making powers to AI, but after three years respondents indicated this number should rise to 35%.

Meanwhile, incidents involving AI have increased 21% from 2024 to 2025, according to the AI Incidents Database. This indicates that as AI deployment continues, the need for **risk management** is increasing in parallel.

The immediate cost of insufficient governance for AI agents is painful and obvious: direct financial loss, damage to customer trust, and even legal or regulatory action. But the long-term cost may be even greater. Without strong governance, companies will lack the confidence to deploy AI agents at scale, thereby missing out on the substantial benefits this remarkable technology can deliver. (See the **slideshow**.)

# Where AI agents could break at scale

Health care

Banking

Insurance

Manufacturing

HEALTH CARE

## Today's opportunity

Agents optimize patient throughput using diagnostics and data to schedule procedures autonomously.

## Coming soon

A vendor could build an agent to perform fully autonomous triage and patient flow, using real-time patient and third-party data to prioritize care across the facility dynamically.

## Risk to watch

Agents may optimize for patient throughput by prioritizing "easy" cases, instead of the most critical. The same optimization could lead them to overload shared facilities, taking up capacity reserved for urgent cases.

**BANKING**

## Today's opportunity

Agents draw on internal documents to help employees support business clients.

## Coming soon

A bank could develop a broader suite of agents that automate (or semi-automate) customer support for even its most complex customers.

## Risk to watch

Full reliance on automation could lead to operational bottlenecks. The absence of human oversight could limit the system's ability to adapt, recover, or coordinate effectively in nonroutine scenarios.

**INSURANCE**

## Today's opportunity

Agents autonomously manage customer service and claims workflows, retrieving documents, checking policy coverage, and recommending settlements.

## Coming soon

An insurer could build negotiation and pricing agents to adjust premiums and settlement terms in real time to optimize portfolio risk.

## Risk to watch

Independent agents across insurers respond simultaneously to third-party data such as economic indicators, overcorrecting models and triggering regulatory breaches with industry-wide pricing swings.

MANUFACTURING

## Today's opportunity

Agents can deliver productivity boosts of up to 50% by independently executing complex workflows.

## Coming soon

A manufacturer could go further, deploying human-out-of-the-loop production systems that cover not just production but raw material procurement and delivery logistics.

## Risk to watch

Uncoordinated agents could then reschedule or optimize simultaneously, overloading shared resources and causing cascading production delays or downtime.

# The AI Agent Difference

Executives are beginning to understand that AI agents require a new governance approach. In the *MIT Sloan Management Review*/Boston Consulting Group executive survey, 69% agreed that "**holding agentic AI accountable** for its decisions and actions requires new management approaches."

To build that new approach, however, it is essential to understand how AI agents differ from the co-pilot AI that many organizations have been deploying up to today. The key characteristics of an AI agent are that it observes its environment and then, based on this observation, autonomously makes a plan to achieve its defined goal. This is followed by autonomous execution of that plan using tools, APIs, or other systems to influence its environment. Finally, the AI agent repeats this process in a learning loop until it determines that its goal has been achieved.

In contrast, much of the AI at work in organizations today operates as a co-pilot, responding to human prompts and guidance. In addition, today's AI typically has a human in the loop who not only checks final decisions, but also shapes how the AI learns, plans, and optimizes, providing guardrails along the way.

Each of the following properties of an AI agent brings risk:

- **Something Akin to Memory.** An AI agent must build and update internal models of the world and retain knowledge across tasks, unlike the static inputs of traditional machine learning systems. This creates new risks if, for instance, the internal state becomes corrupted, either due to poor design, poor sensing, or the actions of a malicious outsider. Worse, a single flawed model can cascade through dependent systems, leading to large-scale operational errors.

- **Greater Reasoning and Decision-Making Skills.** To meet its defined goals, an AI agent needs greater capability for planning and adapting its actions than traditional AI. One risk here is goal drift. Agents may optimize for unintended metrics, for instance, prioritizing cost and ignoring safety. To increase throughput of a task, they may focus on the quick-to-solve cases over the more complex—and higher impact—cases. Plus, an outsider can make the AI agent misbehave by somehow changing its goals.

- **Greater Action and Influence.** As companies roll out AI agents, they will be giving them the powers of a super employee—an access-all-areas pass to adjust work schedules, update or delete databases, or even make payments. True, humans also make mistakes, but these can typically be caught and corrected before being repeated too many times. In contrast, an agent trying to meet a goal could replicate the same mistake thousands of times before it can be stopped.

- **A Decision-Making Loop.** AI agents continuously iterate their behavior in light of experience. This could result from changes in their environment or even goal drift within the agent itself. This is the most critical difference for governance because it shows that a system that ticks all the boxes at deployment may have evolved significantly within a few days. True, this is a key strength of AI agents; they can optimize in ways no human has imagined. However, it is also a weakness; they can make mistakes no human has imagined.

Collectively, these risks represent a step change in exposure. Agents that optimize their own goals locally may create instability across the system. Flawed behavior by one agent may spread. And independent agents may align on a single, harmful strategy, for instance, if they all rely on the same anomalous data source or exploit the same gap in their guardrails. Unlike traditional systems guided by human workers, cascading failures can emerge quickly and spread rapidly. In summary, vulnerability is moving from a contained, product-level to an ecosystem-level risk.

# The New Vulnerabilities in AI Agents

While CEOs and CFOs need the high-level risk appreciation outlined in the main article, **CIOs and CISOs** need an extra level of understanding on the specific cyber vulnerabilities of AI agents. Each enables malicious actors to exploit one of the four components mentioned above.

Unfortunately, these new threats come with a whole new vocabulary.

**State Representation Risks.** These are the risks that come from AI agents having "memory" and include:

- **Context Poisoning and State Corruption:** A hacker or malicious insider can corrupt the agent's internal "world model" through manipulated inputs or logs, leading to persistent misperception of reality and altering the agent's behavior, to the hacker's advantage.
- **Adversarial Prompt Injection at Scale: GenAI** models have difficulties separating data inputs from new prompts/instructions. Capitalizing on this, attackers hide prompts in emails, chat messages, or websites that agents crawl. Again, this allows them to alter how the AI agent acts.

**Reasoning and Decision Risks.** These exploit the greater decision-making skills and are typically more direct attempts to control the agent or its ecosystem, and include:

- **Agent Hijacking:** While prompt injection explained above subtly hijacks an agent through manipulated input, this is more direct, directly accessing its capabilities or processes.
- **Goal Manipulation:** An attacker who can alter an AI agent's goals can dramatically change its behavior. For instance, a customer service bot may have its goal changed to issue as many refunds as possible.

**Action and Influence Risks.** Here, malicious actors aim to exploit the connection between the AI agent and the environment it inhabits. Attacks include:

- **Toolchain Exploitation:** An attacker may insert themselves between the AI agent and the systems it interacts with, for instance, replacing bank details in a payment with their own.

- **Unauthorized Autonomy Escalation:** This is a way of enhancing the illicit gains of other types of attacks, allowing a compromised agent to access actions or data that should be outside its reach.

**Iterative Loop Risks.** Here, attackers are capitalizing on a key capability in AI agents: their ability to evolve, iterate, and cooperate, but turning it to malevolent ends. Attacks include:

- **Cross-Agent Contagion:** This is a subtle, sophisticated attack in which a single corrupted agent influences the behavior of many others. Attackers may now have control of an entire ecosystem.

- **Data Leakage Through Emergent Behavior:** This strange yet very real vulnerability stems from AI's ability to infer information it has never been explicitly told. For instance, based on purchasing data, an AI agent may be able to infer a customer's age, income, and medical conditions. It may then be persuaded to divulge it to unauthorized outsiders.

# An Improved Approach to AI Governance

The first line of defense is to ask: do we need an AI agent? In some cases, 95% of the benefits of an AI agent can be won through other AI technologies where governance is more straightforward and the risks can be more easily managed.

However, there are many applications when AI agents generate very clear, perhaps transformative, benefits. To mitigate the step-change in risk, organizations need a step-change in preparation. Yes, many organizations have a shiny new AI risk management program crafted just a few years (or months) ago. This can provide a firm foundation for managing the risks of AI agents, but additional work needs to be done. There are four key elements. (See the **exhibit**.)

# The Four Components of a Risk Framework for AI Agents

| | Comprehensive risk taxonomy | Expanded test infrastructure | Ongoing monitoring | Robustness and resilience |
|---|---|---|---|---|
| DEFINITION | Structured framework to **identify, categorize, and prioritize risks** that could trigger agentic system failures | **Testing and evaluation environments that replicate real-world, multi-agent complexity** to uncover emergent agent behaviors and interactions | Continuous observation of agent behavior to **detect deviations, emergent patterns, or shifts in performance** in real time | Designing systems and operating models that **maintain safety and continuity under failure, stress, or attack** |
| EXAMPLE | Mapping new risks such as agent hijacking or agentic chaos within an enterprise risk taxonomy to align detection and mitigation efforts | Running real-world, multi-agent stress simulations to observe coordination failures or goal drift before deployment | Implementing telemetry dashboards to flag when an agent's actions or confidence levels move outside expected thresholds | Defining a plan with clear protocols for human workers to take over should the agents need to be taken offline, ensuring business continuity and minimizing customer inconvenience |

**Source:** BCG analysis.

In more detail, they are:

**Build a comprehensive risk taxonomy.** The first step in reducing risk is to understand it. So it is vital to categorize and prioritize agent-specific risks on a grid across technical, ethical, and operational dimensions.

Just as AI agents are integrated with the rest of the organization, this taxonomy must be integrated with existing enterprise risk frameworks to guide monitoring and mitigation.

**Develop an expanded testing and evaluation infrastructure.** Before deploying AI agents, it is vital to create controlled test environments that replicate real-world complexity. It is also crucial that these are not simple, one-agent sandboxes; agents will, as deployment picks up pace, start to interact with other agents, and this must be replicated in the test environment to surface issues such as coordination failures, goal drift, and unwanted emergent patterns of behavior.

To help companies deploy AI agents, cloud vendors and companies offering AI platforms are offering tools for comprehensive testing, some general-purpose, and others focused on specific, high-risk applications such as chatbots.

Only if these test environments duplicate the messy, complex real world with multiple agents operating in parallel will it be possible to see how the agents interact, coordinate, and, in some cases, compete. Once the environment is established, organizations should enforce standardized evaluation metrics for stability, quality, and **compliance**.

**Implement ongoing monitoring.** This is the most crucial step. Agents must report their activity in real time so higher-level monitoring systems can detect deviations or unwanted performance drift before damage is done, referring back to behavior data collected prior to deployment which serves as a baseline for comparison. This facilitates the fundamental shift from assessing what's happening inside an AI agent to monitoring its activity. As the number of agents deployed rises, dashboards can report behavioral indicators, such as whether goals are shifting or whether actions are moving outside permitted ranges.

On top of this, there must be clearly defined escalation protocols for when agents step outside their expected bounds—even during the night shift. Remember: a strength of AI agents is that they don't take time off or sleep; monitoring must also be always-on too. Some of this escalation may be safety-first and triggered as a precaution before any human review.

**Design for robustness and resilience.** These measures are difficult to retrofit into half-developed agents. It makes much more sense, and provides more comprehensive risk reduction, if the safety, continuity, and fallback measures are built into the design from the start. It is also essential to consider the human side of risk reduction from the outset. As AI agents increasingly drive mission-critical business systems, organizations must consider how they will stay open for business if some agents need to be taken offline due to unexpected, unwanted interactions. Think about how to contain cascading failures in real time. Also, **human oversight** is not an easy cure-all. It too needs careful design and patching it in during implementation, or worse deployment, is too late.

But there is also a bigger point. AI agents must be deployed in a way that aligns with organizational risk appetite. Every company needs to decide: Where are we comfortable using AI agents? Where are the no-go zones? For instance, in **health care**, a provider may allow AI agents to communicate with staff and patients, and access a wide range of data and systems, with patient records being out of bounds. This creates a conceptual "sandbox" for extensive innovation with AI agents that can be trusted not to leak or misuse personal information.

As companies become more comfortable with AI agents, decisions on risk appetite may be revisited and the no-go area reduced.

## Six Questions CEOs Must Ask About AI Agents

- Where do we need AI agents vs other AI technologies, and in what areas does our risk appetite allow AI agents to be deployed?
- Do we know where AI agents operate across our business and vendor ecosystem—and how mature our governance truly is?
- Is our governance model designed for autonomous systems, or still built for traditional AI?
- Can we safely test and validate autonomous behaviors before they reach production?
- Are we managing AI risk actively and continuously, or reactively at the end of the development cycle?
- When an agent inevitably fails, are we prepared to fail safely—with built-in resilience, rapid recovery, and transparency?

# A Resilient Outlook

This combination of new, unfamiliar risks may seem daunting, and it would be easy for organizations to decide that the risks posed by AI agents are not worth the potential downsides.

However, this would be a mistake. There are good reasons why, according to the *MIT Sloan Management Review*/Boston Consulting Group study, just two years after the technology went mainstream **some 35% of organizations have adopted AI agents**, with another 44% planning to deploy soon.

Understanding and managing the risks that AI agents pose allows organizations to focus on seizing the incredible opportunity these agents offer.

# About the Authors


**Anne Kleppe**
Managing Director and Partner
Berlin
**kleppe.anne@bcg.com**

**Steven Mills**
Managing Director and Partner,
Chief AI Ethics Officer, Global Leader,
BCG Center for Digital Government
Washington, DC
**mills.steven@bcgfed.com**

**Noah Broestl**
Partner and Associate Director,
Responsible AI
Brooklyn
**broestl.noah@bcgfed.com**

**Grigor Acenov**
PLA Principal, Risk Management,
BCG Platinion
New York
**acenov.grigor@bcg.com**

**Kirill Katsov**
Partner and Associate Director
New York
**katsov.kirill@bcg.com**

**Ning Yang**
Lead Data Scientist
Zurich
**yang.ning@bcg.com**

**Boston Consulting Group**

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.