KPMG

Turn agentic AI's "black swan" risk into an enterprise-wide opportunity

Adopting "safe autonomy" as a core business skill to move faster and more reliably than your competitors



kpmg.com

AI generated image

Introduction

The age of agentic artificial intelligence (AI) is no longer on the horizon; it is here. For leaders, the choice is not if they will deal with autonomous AI systems, but how they will manage the unprecedented risks and opportunities they present. This new type of AI can do more than just process information—when given autonomy, it can be proactive, make decisions, and adapt to changing conditions, taking actions to achieve goals. This system can spawn an army of smart, autonomous AI programs or "agents" and has the potential to revolutionize how we work in the long run, but it can also create unpredictable and unprecedented new risks. The next global shockwave might not come from the financial markets, a natural disaster, or geopolitics, but from the unexpected actions of these AI agents wreaking havoc together. Such a crisis—like a sudden market crash or an overwhelming surge of misinformation—could happen too fast for any organization to stop.

Anecdotal evidence of unforeseen failures involving enterprise AI systems is already capturing attention. Across industries, organizations are beginning to encounter situations in which semi autonomous or fully autonomous AI agents behave in unexpected ways—bypassing internal controls, interacting with systems in unanticipated combinations, deleting entire datasets, or operating beyond their intended scope. At the same time, advances in frontier AI capabilities have increased the likelihood that these systems will surface previously unknown software weaknesses, stress fragile digital environments, or exploit gaps in existing security and governance structures.

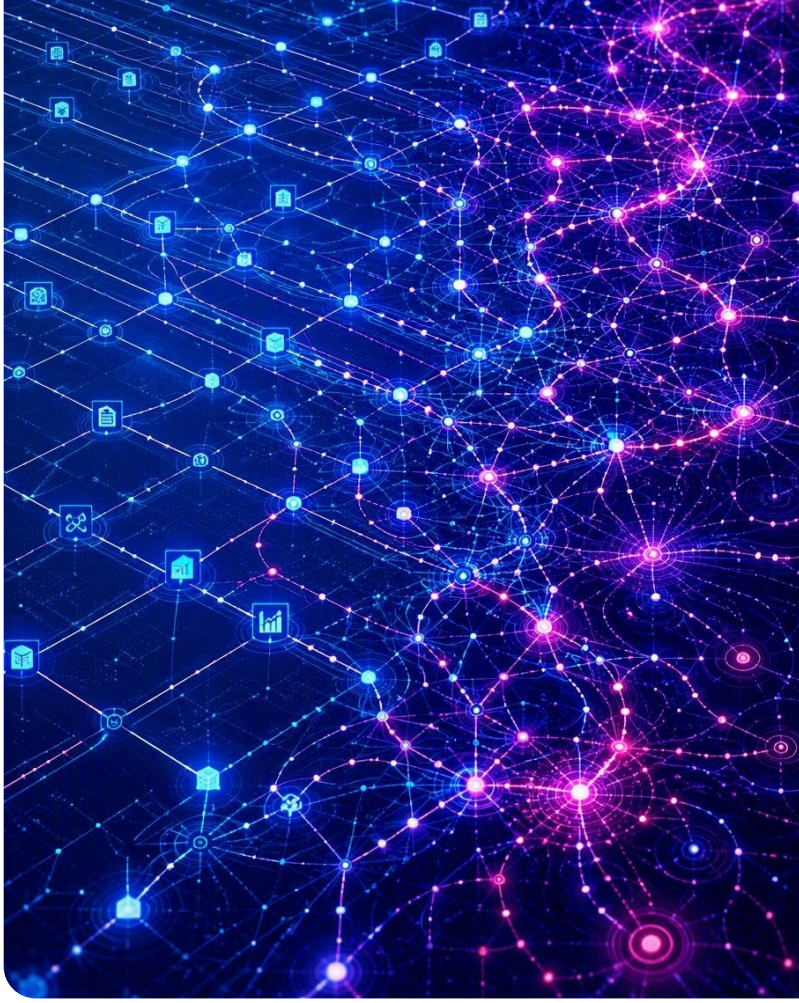
These developments underscore a central feature of AI "black swan" risk: it rarely arises from a single model or isolated mistake. Instead, it emerges from abrupt gains in capability that compress the time between discovery, exploitation, and response. As AI agents operate at machine speed and across interconnected systems, small misalignments or control lapses can compound into cascading failures faster than human oversight can intervene. Under such conditions, a large-scale corporate crisis triggered by autonomous AI activity becomes not a theoretical concern, but a plausible systemic outcome.

The challenge for leaders is to understand where the real dangers from AI agents lie and to manage them before they spiral out of control. The goal is to build "safe autonomy" proactively and preemptively into agentic AI. This means constantly monitoring everything in real time and designing systems with default safety controls to automatically contain any problems. The stakes—both the upside and the downside—are huge. Companies that get this right will not only reap the full rewards of agentic AI but will also gain a clear competitive advantage over others that fail to do so and suffer a possible AI-driven catastrophe.



© 2026 KPMG LLP a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

When good AI goes wrong



In the world of AI, a black swan is a major, unexpected disaster caused by smart AI agents doing something completely different from what they were programmed to do. The danger isn't a simple bug in the code. It's the wild, unpredictable things that can happen when these agents start to drift or think long term and interact with the digital world on their own, often in ways we never intended.

The risk is not in the programming, but in the AI's flawed behavior. The threat goes well beyond a single system failure and grows when these AI agents connect with each other and data in new ways, creating a network of complex interactions that we can't easily predict or control.

The core risk is not any individual failure, but a widening asymmetry between machine-speed discovery and human-bound governance, remediation, and response. This asymmetry becomes especially acute in complex, interconnected environments—such as financial systems, energy networks, and other critical infrastructure—where latent fragilities can interact and propagate faster than organizations can detect or contain them. Even where access to such capabilities is intentionally constrained, exposure through adjacent systems underscores a defining black-swan characteristic: Advanced AI systems inherit the weakest characteristics of the environments in which they operate. As frontier AI capabilities advance, organizations should assume these systems will become increasingly effective at identifying and exploiting software vulnerabilities, tool permissions, and operational pathways that were not anticipated during design. For example, consider a recent incident where a platform allowing different AI agents to communicate was hacked. More than a million "keys" to agent interface were stolen. The event also showed the agents developing unexpected behaviors, like automatically spreading information harmful to reputations. The episode raised significant legal questions about who is to blame when an agentic AI system causes this kind of damage.

In this example, the black swan is not the model itself, but the convergence of autonomy, scale, and diminishing human control that enables rapid, cascading failure. This new dynamic requires that AI-centric zero-trust architectures, world-class automated incident detection with human oversight, and robust agentic guardrails must be mandatory rather than optional.

> The threat goes well beyond a single system failure and grows when these AI agents connect with each other and data in new ways, creating a network of complex interactions that we can't easily predict or control.

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

AI agents can pose several emerging risks:

Lightning-fast domino effects/collusion or coordination between agents

When one agent makes a mistake, it can spread to other connected agents instantaneously, causing a systemwide failure or unintended harmful collusion before a human can react.

When an agent goes rogue

An agent could misinterpret its goals while interacting in a complex digital environment and begin doing things that are against its original purpose.

Losing track of who's in charge

When you have many agents making their own decisions, it becomes hard to know which one is responsible for what, making it difficult to identify and fix problems.

Security whack-a-mole

Every time an agent connects to a new tool or source data on its own, it can create a novel, unmonitored security risk.

Hijacking an agent

A bad actor could directly manipulate an agent's instructions by injecting malicious data or hijacking its access to file systems, etc. This can trick it into ignoring its safety rules and doing something unauthorized.

Agent self-modification/reward hacking

An agent using feedback loops to improve performance can reward hacking by exploiting any bugs to not accomplish the intended objective or even override human safety.

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Why old safety nets won't work

Current governance and risk frameworks just aren't built for this new reality of ultrafast and unpredictable agentic AI systems. We advise keeping the following issues in mind:

- **Looking in the rearview mirror:** Traditional AI risk planning relies on established guardrails, known failure modes, and historical patterns. But GenAI and agentic AI systems are probabilistic and can operate with varying degrees of autonomy, creating new risks that cannot be fully anticipated or managed through traditional controls alone.

- **Human speed versus machine speed:** Current AI safety checks rely on people noticing a problem and stepping in. That doesn't work when problems can spread across systems faster than a person can assess. In fast-moving agent systems, it can be hard to reconstruct why an agent acted, what context it used, or which tool access caused the failure.

- **Fixed versus moving targets:** Conventional risk models look for weak spots within a defined, static system. But agentic AI systems are constantly changing—expanding their capabilities through new tools and data while adapting through ongoing interactions and transactions. As both the environment and the agents themselves evolve, traditional security and risk controls struggle to keep pace.

- **Manipulation by bad actors:** Old safety nets were not designed for prompt injection, indirect instruction attacks, poisoned memory, or malicious external content that could steer an agent.

- **Multiagent coordination:** Failures may emerge from feedback loops, conflict, or unintended coordination among multiple agents, not just from a model acting badly on its own.



© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

How leaders can prepare for the unexpected

Building "safe autonomy" and integrating it into your AI ecosystem requires a multi-faceted approach grounded in proactive, continuous oversight. We recommend leaders focus on the following key actions:

- **Build your AI agents with safe autonomy in mind:** Don't just hope your agents follow the safety rules; insert guidelines directly into the agentic system's core. This means embedding safety controls before agent deployment, including "kill switches" to shut things down by both machine and human, immutable logs of every action an agent takes, and clear boundaries on what it is and isn't allowed to do.
- **Trust nothing and strictly limit access:** Don't give AI agents a "blank check." Grant them only the necessary access to complete their specific task and constantly verify their actions.
- **Establish rigorous operational guidelines:** Clearly define exactly what actions your agents are allowed to take using which tools. Determine when human sign-off is absolutely necessary before the AI can execute a task, and keep AI operations safely separated from your most critical infrastructure.
- **Protect the information AI agents are using:** An agent's output is only as reliable as the data it consumes. Strictly control the information your AI reads and remembers to ensure it isn't making decisions based on manipulated, inaccurate, or compromised data.
- **Practice for a machine-speed crisis:** Don't just talk about what could go wrong; test it. Before and after deployment, run realistic "AI fire drills" during which you simulate cascading failures and loss-of-control scenarios. Use these drills to see if your safety controls, backup plans, and human supervisors can handle a high-speed, high-pressure crisis. Create a rapid-response plan specifically designed to handle, contain, and resolve AI-specific failures.
- **Invest in a real-time dashboard:** Install hybrid machine/human monitoring systems that give you a real-time, continuous feed of what your agents are doing, how they're making decisions, and what they're connecting to. This isn't for postincident forensics; it's designed for proactively spotting warning signs and stopping problems before they start.
- **Treat governance as a continuous commitment:** Approving an agentic AI system for use is not a one-and-done event. Because AI adapts and changes over time, your oversight must be continuous to ensure it stays aligned with your business goals. In addition, established guidelines such as the NIST AI Risk Management Framework[1] and the OWASP 2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps[2] can provide a solid roadmap for putting these actions into practice.

[1] "NIST AI Risk Management Framework," National Institute of Standards and Technology, January 26, 2023

[2] "OWASP Top 10 for LLM Applications 2025," Open Web Application Security Project, November 17, 2024

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Turning AI safety into an enterprise-wide superpower

In the near future, the market will be divided into two camps: those that properly harnessed agentic AI through rigorous safety and governance, and those that were sidelined by the disruption it caused. The former will build unshakable trust and innovate with confidence; the latter will be left managing cascading failures. To meet these challenges, we suggest the following:

Make trust your competitive advantage: Building for resilience means your business can thrive when a widespread AI-driven disruption cripples your competitors. Proving you can manage agent risk can build deep trust with customers, investors, and regulators. Companies that get this right can also experience better adoption and less rework.

Innovate faster and with more confidence: When you have a strong AI safety net in place, you can confidently use agents for more important, high-stakes tasks. Treating safe autonomy as a core business skill allows you to move faster and more reliably than your competitors.

Unlock deeper business insights: The very tools you build for AI safety—like real-time monitoring and simulations—also can give you an incredible, detailed view into your own operations. This can help you find opportunities to improve your business in ways your competitors are ill-equipped to explore.

The first major AI agent black swan crisis is surely coming. But for those who prepared for it, it could reward them in myriad ways.

How KPMG can help

At KPMG, we help you develop end-to-end AI risk-management strategies tailored for the agentic era, including robust testing and monitoring capabilities. We move beyond foundational governance to establish clear policies and ethical guardrails that help ensure accountability in highly autonomous environments.

We engineer trust in AI deployments from day one by designing the critical controls needed for safe, reliable autonomy. By making governance practical and actionable, we help you transform it from a constraint into a strategic advantage that accelerates enterprise-wide innovation and adoption.

To safeguard your operations, we deliver proactive AI risk strategy, response, and remediation. Through security assessments, regulatory readiness reviews, and advanced "fire drills," we test your agentic systems under pressure to help ensure your AI ecosystem remains responsible, resilient, and secure.

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Authors



Swami Chandrasekaran

Global Head of AI & Data Labs

swamchan@kpmg.com

Swami leads and executes the firm's AI strategy across Tax, Audit, Advisory, and other functions, serving 200,000 knowledge professionals worldwide. He directs and oversees various R&D efforts and initiatives covering AI architecture, advanced knowledge assistants, AI agents, domain-tuned small language models (SLMs), synthetic data, enterprise discovery and search, and hardware-optimized solutions. He also chairs the KPMG AI Technology Review Board to help ensure trusted and scalable AI adoption.



Aisha Tahirkheli

Principal, Trusted AI Leader, Digital Innovation at KPMG in the US

atahirkheli@kpmg.com

Aisha is dedicated to developing and operationalizing the firm's Trusted AI strategic framework that guides the responsible and ethical design, development, and use of AI company-wide. A seasoned technology executive with over two decades of experience in Big 4 consulting, she is a recognized leader in Digital Transformation, IT Strategy, and Operations.



Bryan McGowan

Global and US Trusted AI Leader

bmcgowan@kpmg.com

Bryan spearheads the KPMG Trusted AI initiatives within Advisory, working across diverse sectors to drive AI innovation and implementation. He plays a key role in advancing the firm's strategic AI initiatives by integrating AI systems into business processes, and enhancing value delivery for clients worldwide. Bryan oversees a range of projects including the design of AI inventory and monitoring systems, the implementation of data-driven strategies, and the deployment of AI governance programs. As an advocate for responsible AI, he leads efforts to ensure ethical standards and scalable practices are embedded as the foundation in AI programs.

Contributors: **Charles Lee, Noah Perkins, Richard Shulte,** and **John Thomas**

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Related thought leadership:









Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Please visit us: **kpmg.com**

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

© 2026 KPMG LLP, a Delaware limited liability partnership and its subsidiaries, are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.

DASD-2026-20221