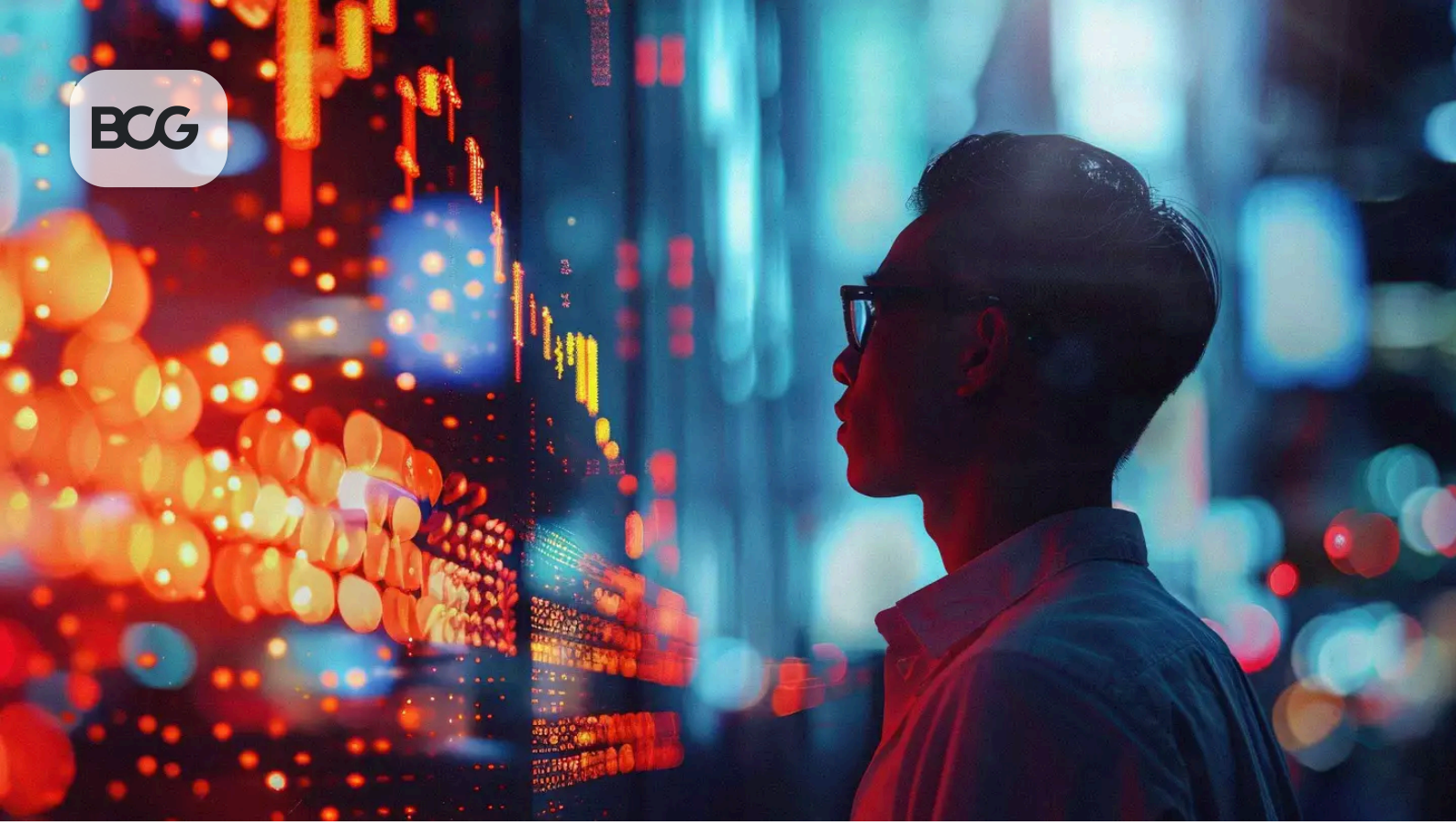

RESPONSIBLE AI

AI Risk Management Needs a Better Model

By Steven Mills, Jeanne Kwong Bickford, Kirill Katsov, and Grigor Acenov

ARTICLE MARCH 26, 2026 8 MIN READ

As the upsides and downsides of AI systems come into focus, executives have moved quickly to launch risk and quality management programs. While effective, these programs have inadvertently created friction by implementing one-size-fits-all processes and duplicative reviews that slow innovation and adoption.

Companies need a new approach to managing AI quality and risk—one that is fast and fluid for familiar uses of AI but thorough and deep for novel and unproven uses, including self-executing agents. This approach removes the bureaucracy that can slow innovation yet ensures deep review for uses that need it most. Companies also must ensure their approach is future-proof, able to continuously evolve as agents gain increasingly sophisticated, autonomous capabilities.

To build this new approach, companies can borrow from two familiar concepts:

- The first is the triage system used by electric utility providers, in which the most serious and widespread outages receive the most attention. For AI cases, that means thorough review, specialized testing, and the application of strong guardrails and risk mitigations.
- The second is the saying "Don't reinvent the wheel." After companies develop proven playbooks and tools to manage the risks of a specific AI activity, they can re-use them for similar AI use cases. The novelty of AI and AI agents seemingly prevents this commonsense approach from taking hold at many companies.

A smart governance approach to AI does more than manage risks more effectively. It improves the overall pace of innovation and quality of AI deployments.

What's Not Working

Many governance programs were designed for yesterday's AI. They assumed a small number of deployments managed centrally, released carefully, and governed by a standard process.

Those assumptions are being blown to bits by the rise of AI agents, vibe coding, and other developments. At many companies, the AI portfolio has expanded from a handful of models to hundreds of systems and tools. Product teams, functional teams, frontline employees, and centralized tech teams are all contributing to growth.

Traditional governance models are collapsing under the volume. When every AI use case is treated the same regardless of maturity, exposure, or business need, progress slows. Low-risk requests rob high-risk initiatives of the deep attention they deserve.

Teams and individuals hoping for fast approval face long, bureaucratic processes. Several functions—risk, legal, security, engineering, and the business—review the same submission, often reaching different conclusions. Good ideas are abandoned out of frustration or, worse, teams route around the governance bottleneck, creating shadow AI.

The current system creates a dilemma for organizations:

- Move too slowly and fall behind in the market, miss internal operational gains, and frustrate your employees.

- Move too fast and increase the odds of product failures, compliance breaches, and reputational damage.

A Better Way

The resolution to this dilemma is a new approach: different AI uses have different risk profiles and require different levels of review and mitigation. The insights that teams have gained from prior use cases are not lost. They are codified in a knowledge base of risks and successful guardrails and mitigations.

This knowledge base has two primary benefits:

- New teams can turn to this playbook to learn how identical or similar risks were successfully managed in the past.
- It can serve as a central control point to build an inventory of AI uses. This critical inventory will allow companies to understand their AI exposure and supply chain risks if new vulnerabilities are discovered or regulatory actions are imposed.

In this approach, most requests are handled swiftly by applying proven guardrails and mitigations. But high-stakes, novel, or unproven applications, especially those involving agents, receive extra attention and expertise.

Risk management becomes an ongoing organizational capability built around speed for most cases and thorough diligence for novel uses where the risk is material. Rather than being a pesky, check-the-box activity, risk management promotes innovation and quality. It adds value rather than friction.

Managing AI Risks in Real Time

This improved approach reduces business friction by moving from ad hoc, inefficient processes to a streamlined, structured, and well-understood approach. A sponsoring team answers a short

series of questions about the proposed use of AI, and each question is explicitly tied to a potential risk.

Collectively, the answers to these questions accurately assess the inherent risk of the AI use. The system then automatically routes the request based on risk level (how big is the impact if things go wrong?), novelty (have we seen a similar use before?), and readiness (do we already have guardrails in place?). Applications will fall into one of four tiers. (See the exhibit.)

A Four-Tiered Triage Approach to Manage AI Risks

Tier	Description	Criteria	Oversight required	Share of applications
Self-Service	**Well-known use cases** that leverage current technology and involve well-understood use patterns	Low risk	Delivery teams follow standard best practices for guardrails, controls, and system design	~75–90%
Trust but Verify	**High-risk use cases** but based on known tech patterns, established use cases, and proven mitigations	High risk and known use	No formal approval needed; independent oversight as needed during design, deployment, and use	~5–20%
Strategic Review	**High-impact use cases** with significant and unfamiliar risk exposure	High risk and novel risk	Formal approval required; independent oversight needed for design, deployment, and use	~5–15%
Prohibited	**Use cases that run counter** to the organization's risk appetite, values, or regulatory requirements	Prohibited	Reject	<1%

Source: BCG experience.

Self-Service. This tier covers common uses with low inherent risk. The team can move forward following standard best practices for AI quality and risk mitigation without waiting for approval. The project is tracked as part of the AI inventory to enable effective oversight during execution. These applications can be handled in hours. They could conceivably cover at least 75% of requests.

Imagine a team that wants to run an AI assistant that summarizes internal meetings and then drafts follow-up emails for the organizer to send. The overall risk of this system is low, and it can proceed quickly without deep review.

Trust but Verify. Included here are uses with elevated but well-understood risks with proven mitigations. Similar applications have been reviewed in the past. Although the context may be slightly different, these applications can proceed using the established mitigations.

These mitigations may include well-tested models, prompts, and guardrails built on enterprise AI platforms or other tools. The review can be conducted quickly, focusing on the updated context. These applications can be handled in days. Up to 20% of applications may fall in this lane.

Imagine a company has already deployed an AI chatbot for customer service using an approved enterprise LLM, standard guardrails, and human review for edge cases. Now the business wants to extend the same chatbot to a new product line and add a small capability—like auto-drafting

call summaries for agents. The model, architecture, and data-handling approach are familiar, but the context is slightly expanded.

Strategic Review. This tier focuses on high risks and novel risks for the organization. This level of review requires experts with deep experience in AI technology, testing, and risk mitigation. A cross-functional team of reviewers would need to thoroughly map and assess technical, business, legal, information security, and other risks. For each case, the review team would need to determine mitigations that reduce risk to acceptable levels.

Depending on the risk tolerance of the organization, these reviews would be reserved for 5% to 15% of applications. Initially, more applications would likely go through this process, but the numbers should decline over time as companies develop proven guardrails and mitigations, allowing more applications to belong in the trust-but-verify tier.

Imagine an AI assistant that drafts credit memos for small-business lending by summarizing financials and proposing risk factors and terms. The review team would need to thoroughly test and evaluate the system, document it, show how customer data is protected, and define guardrails and remediation steps if the system produces incorrect or harmful recommendations.

Prohibited. This tier encompasses potential uses that present an unacceptable downside risk, running counter to an organization's risk appetite, values, or regulatory requirements in specific jurisdictions. Prohibited AI uses are often documented in advance by an organization as part of their AI code of conduct or equivalent policy.

Consider a proposal for an autonomous agent that can change customer account settings or initiate transactions without human approval, using a newly adopted model and drawing on external data sources. The potential for customer harm, regulatory exposure, or reputational damage and the lack of effective mitigations create unacceptable risk for most organizations.

To scale AI and AI agents, companies need to develop an efficient and effective approach to risk and quality. The following actions will lead to a smart governance approach that balances speed and safety:

- **Standardize intake.** Create an online form with a streamlined set of questions directly tied to risks for teams to fill out for new uses of AI.
- **Route applications.** Automatically steer applications into four review lanes based on prior experience, inherent risk, and novelty.
- **Create a playbook.** The mitigations that reduce specific risks are codified so that new teams can learn from what worked in the past, and more applications can be fast-tracked for approval.

- **Clarify decision rights.** Ensure clear roles and responsibilities so that the review process is efficient and predictable.

Done well, this approach speeds innovation, reduces shadow AI and duplication, and becomes a source of value creation rather than frustration and missed opportunity.

Authors

Steven Mills

Managing Director & Partner, Chief AI Ethics Officer, Global Leader, BCG Center for Digital Government
Washington, DC

Jeanne Kwong Bickford

Managing Director & Senior Partner
New York

Kirill Katsov

Partner & Associate Director, Risk & Analytics in FI
New York

Grigor Acenov

PLA Principal, Risk Management, BCG Platinion
New York

ABOUT 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.