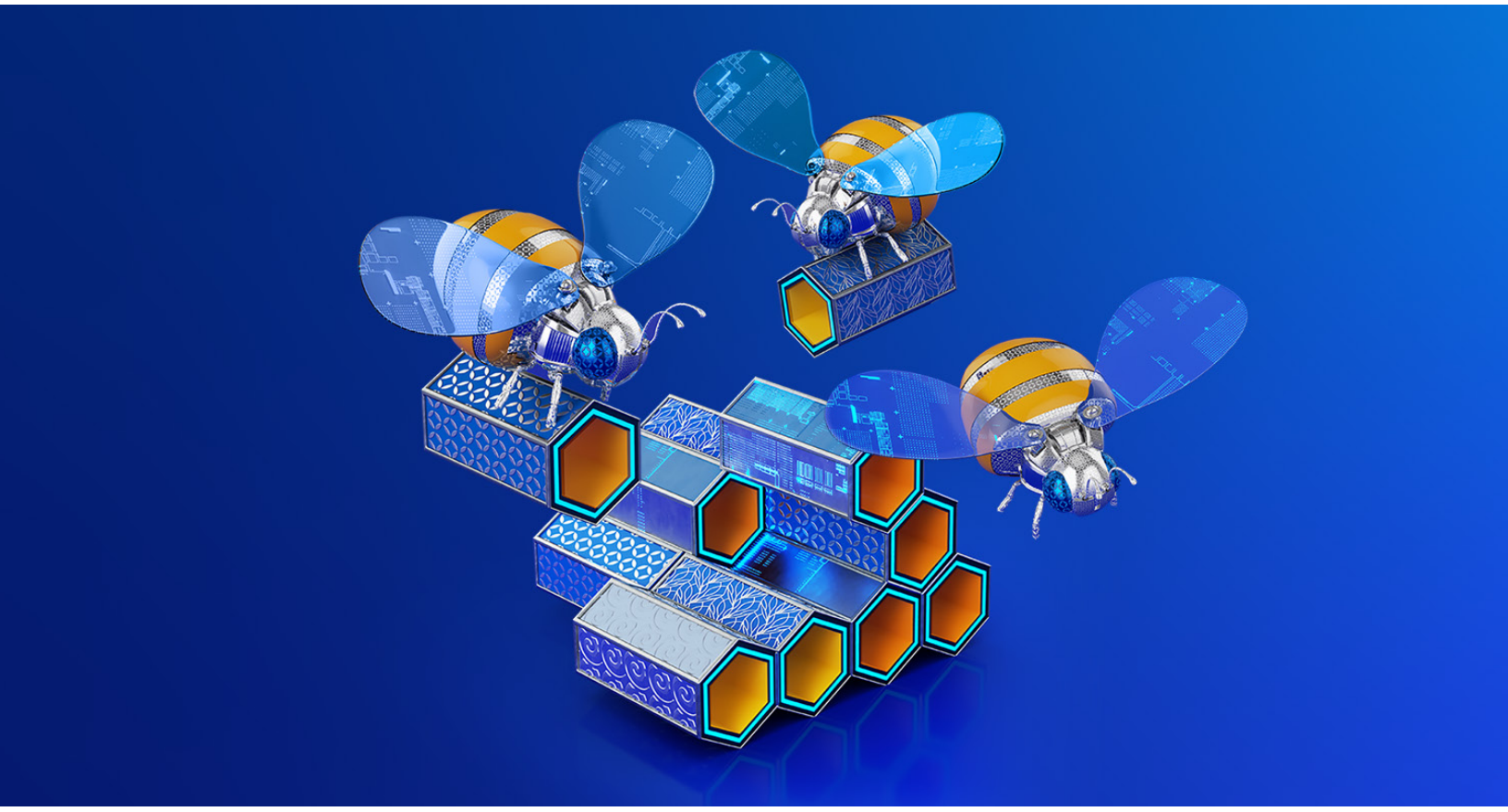
McKinsey & Company | QuantumBlack AI by McKinsey

Implementing generative AI with speed and safety

Generative AI poses both risks and opportunities. Here's a road map to mitigate the former while moving to capture the latter from day one.

This article is a collaborative effort by Oliver Bevan, Michael Chui, Ida Kristensen, Brittany Presten, and Lareina Yee, representing views from McKinsey's Risk & Resilience Practice and QuantumBlack, AI by McKinsey.



March 2024

Generative AI (gen AI) presents a once-in-a-generation opportunity for companies, with the potential for transformative impact across innovation, growth, and productivity. The technology can now produce credible software code, text, speech, high-fidelity images, and interactive videos. It has identified the potential for millions of new materials through crystal structures and even developed molecular models that may serve as the base for finding cures for previously untreated diseases.

McKinsey research has estimated that gen AI has the potential to add up to $4.4 trillion in economic value to the global economy while enhancing the impact of all AI by 15 to 40 percent.[1] While many corporate leaders are determined to capture this value, there's a growing recognition that gen AI opportunities are accompanied by significant risks. In a recent flash survey of more than 100 organizations with more than $50 million in annual revenue, McKinsey finds that 63 percent of respondents characterize the implementation of gen AI as a "high" or "very high" priority.[2] Yet 91 percent of these respondents don't feel "very prepared" to do so in a responsible manner.

That unease is understandable. The risks associated with gen AI range from inaccurate outputs and biases embedded in the underlying training data to the potential for large-scale misinformation and malicious influence on politics and personal well-being. There are also broader debates on both the possibility and desirability of developing AI in general. These issues could undermine the judicious deployment of gen AI, potentially leading companies to pause experimentation until the risks are better understood—or even deprioritize the technology because of concerns over an inability to manage the novelty and complexity of these issues.

However, by adapting proven risk management approaches to gen AI, it's possible to move responsibly and with good pace to capture the value of the technology. Doing so will also allow companies to operate effectively while the regulatory environment around AI continues to evolve, such as with President Biden's executive order regarding gen AI development and use and the EU AI Act (see sidebar, "The United States moves to regulate AI"). In addition, most organizations are likely to see the use of gen AI increase "inbound" threats (risks likely to affect

The United States moves to regulate AI

On October 30, 2023, the Biden administration released a long-awaited executive order aimed at addressing concerns related to AI development in economic, national-security, and social domains. The order establishes principles, tasks federal agencies with AI-testing methods, codifies government oversight of private AI development, and outlines AI's impact on national security and foreign policy:

- *Holistic AI governance.* The order establishes a comprehensive framework for AI governance, emphasizing ethics, safety, and security. It addresses the importance of responsible innovation, collaboration, and competition in the AI industry.
- *Private sector accountability.* The order mandates that private companies involved in AI adhere to industry standards, report on compliance, and implement best practices. This includes meeting specific guidelines on transparency and accountability, especially for dual-use foundation models and large-scale computing clusters.
- *Cross-sector impact.* The order addresses various sectors affected by AI, including critical infrastructure, cybersecurity, education, healthcare, national security, and transportation. It promotes interagency collaboration to integrate AI responsibly and securely across these sectors, aligning government and industry efforts for societal benefit.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[2] Unpublished data from McKinsey survey results.

organizations regardless of whether they deploy gen AI), particularly in fraud and cyber domains (early indications are that gen AI will be able to defeat standard antifraud biometric checks[3]). Building fit-for-purpose risk management will help guard against these threats.

In practical terms, enterprises looking to address gen AI risk should take the following four steps:

1. Launch a sprint to understand the risk of inbound exposures related to gen AI.

2. Develop a comprehensive view of the materiality of gen-AI-related risks across domains and use cases, and build a range of options (including both technical and nontechnical measures) to manage risks.

3. Establish a governance structure that balances expertise and oversight with an ability to support rapid decision making, adapting existing structures whenever possible.

4. Embed the governance structure in an operating model that draws on expertise across the organization and includes appropriate training for end users.

The specifics of how to implement these steps and the degree of change required to make them effective will vary with an organization's gen AI aspirations and nature. For instance, it could be looking to be a *maker* of the foundation models, a *shaper* that customizes and scales foundation models, or a *taker* that adopts foundation models through off-the-shelf applications with little or no customization (for example, standard office productivity software).[4]

This article provides a blueprint for developing an approach to implementing gen AI responsibly. Following these steps helps organizations move quickly to scale the technology and capture its benefits while minimizing their exposure to the potential downsides.

Understanding and responding to inbound risks

In our experience, including through building McKinsey's own gen AI application, gen-AI-related risks can be captured in eight main categories (Exhibit 1). These categories consider both inbound risks and risks that directly result from the adoption of gen AI tools and applications. Every company should develop some version of this core taxonomy to support understanding and communication on the risks arising from the implementation of gen AI.

Most organizations are likely to see the use of gen AI increase 'inbound' threats, particularly in fraud and cyber domains.

[3] *Security Intelligence*, "AI may soon defeat biometric security, even facial recognition software," blog entry by Mike Elgan, January 31, 2019.
[4] For more, see "Technology's generational moment with generative AI: A CIO and CTO guide," McKinsey, July 11, 2023.

Exhibit 1

Half of eight basic categories of generative AI risk apply to all organizations regardless of their deployment of related use cases.

Risk category	Description	Inbound	Gen AI[1] adoption
Impaired fairness	Algorithmic bias resulting from unrepresentative training data or model performance or misrepresentation of AI-generated content as human created		✓
Intellectual property (IP) infringement	Infringement on copyrighted or otherwise legally protected materials, inadvertent leakage of IP into public domain, or both	✓	✓
Data privacy and quality	Unauthorized use or disclosure of personal or sensitive information or use of incomplete or inaccurate data for model training		✓
Malicious use	Malicious or harmful AI-generated content (eg, falsehoods/deepfakes, scams/phishing, hate speech)	✓	✓
Security threats	Vulnerabilities in gen AI systems (eg, payload splitting to bypass safety filters, manipulability of open-source models)	✓	✓
Performance and "explainability"	Inability to explain model outputs or model inaccuracies appropriately (eg, factually incorrect or outdated answers, hallucinations)		✓
Strategic	Risk of noncompliance with standards or regulations, societal risk, and reputational risk		✓
Third party	Risks associated with use of third-party AI tools (eg, proprietary data being used by public models)	✓	✓

[1]Generative AI.

McKinsey & Company

Deciding how to respond to inbound risks is a focus for many executive teams and boards. This decision should serve as a foundation for how an organization communicates about gen AI to its employees and stakeholders. It should also inform the approach to use cases.

We see four primary sources of inbound risk from the adoption of gen AI:

- security threats, resulting from the increased volume and sophistication of attacks from gen-AI-enabled malware
- third-party risk, resulting from challenges in understanding where and how third parties may be deploying gen AI, creating potential unknown exposures
- malicious use, resulting from the potential for bad actors to create compelling deepfakes of company representatives or branding that result in significant reputational damage
- intellectual property (IP) infringement, resulting from IP (such as images, music, and text) being scraped into training engines for underlying large language models and made accessible to anyone using the technology

Most organizations will benefit from a focused sprint to investigate how gen AI is changing their external environment, with two primary objectives. The first is to understand potential exposures to inbound risks, anchored in the organization's risk profile (for example, how many third parties have access to sensitive or confidential data that need to be restricted from training external gen AI models). The second objective is to understand the maturity and readiness of the control environment—the technical and nontechnical capabilities the organization has in place to prevent, detect, and ultimately respond to inbound risks. These include cyber and fraud defenses, third-party diligence to identify where critical third parties may be deploying gen AI, and the ability to limit the scraping of company IP by engines used to train large language models.

The outcome of these efforts should be an understanding of where the organization faces the largest potential inbound exposures, as well as the maturity and readiness of its current defense system. Having conducted this exercise, the organization should have a clear road map of where to harden defenses and what the potential ROI from these efforts would be in potential risk mitigation.

Given the evolving nature of the technology underlying gen AI and its applications, organizations will need to repeat the effort to identify their exposure with some regularity. For most organizations, refreshing this exercise at least semiannually will be important until the pace of change has moderated and the control environments and defenses have matured.

Tethering Prometheus: Managing the risks produced by gen AI adoption

Organizations with ambitions to deploy gen AI will need to undertake additional, ongoing efforts to understand and manage the risks of the technology's adoption. This will likely require an investment of time and resources and a shift in ways of working. Yet it's essential if organizations are to achieve long-term, sustainable, and transformative benefits from gen AI. Missteps and failures can erode the confidence of executives, employees, and customers and trigger scaling back in the level of ambition to ultrasafe use cases that generate limited risk but are also unlikely to capitalize on the technology's true potential.

Organizations looking to deploy high-potential use cases for gen AI to drive productivity and innovation; provide better, more consistent customer service; and boost creativity in marketing and sales must address the challenge of responsible implementation. These use cases have varying risk profiles, reflecting both the nature of the technology itself and company-specific context concerning the specifics of the use case (for example, deployment of a gen AI chatbot to certain at-risk populations has a very different risk profile from that of a B2B deployment) (Exhibit 2).

Exhibit 2

Different generative AI use cases are associated with different kinds of risk.

✓ Primary risk

Generative AI use case	Impaired fairness	IP[1] infringement	Data privacy and quality	Malicious use	Security threats	Performance and 'explainability'	Strategic
Customer journeys *(eg, chatbots for customer services)*	✓		✓			✓	✓
Concision *(eg, generating content summaries)*	✓	✓				✓	
Coding *(eg, generating or debugging code)*		✓		✓	✓	✓	
Creative content *(eg, developing marketing content)*	✓	✓		✓		✓	

[1]Intellectual property.

McKinsey & Company

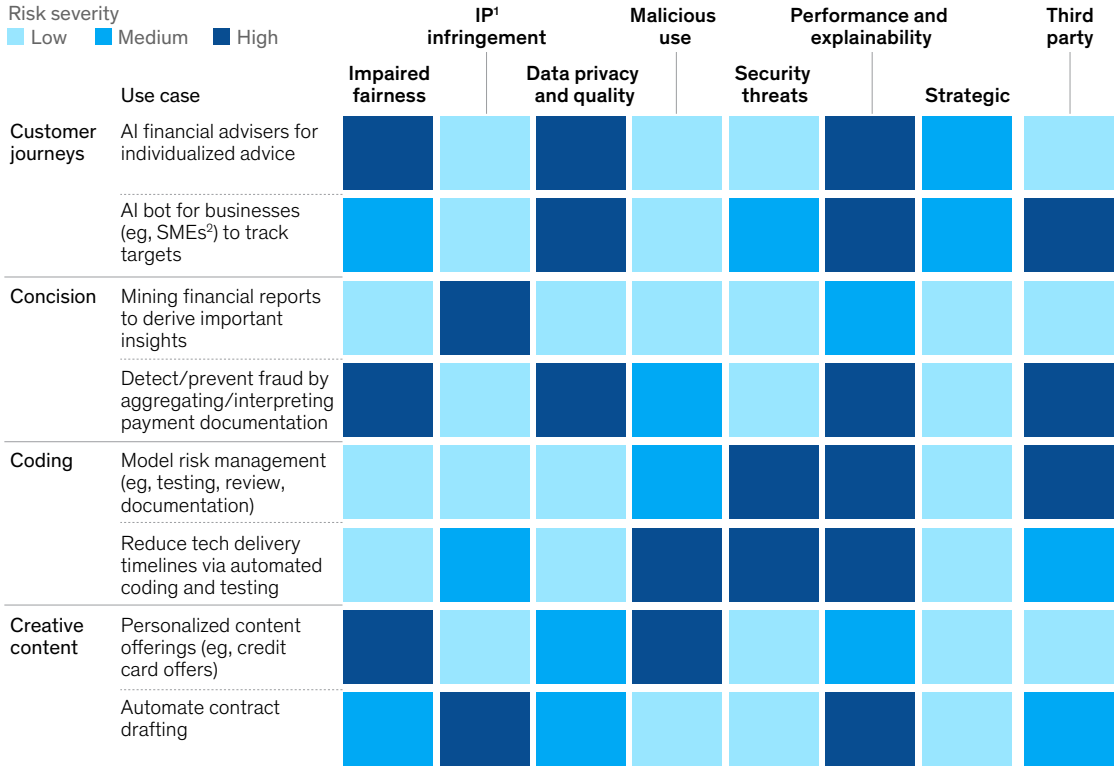
Identify risks across use cases

The essential starting point for organizations deploying gen AI use cases is to map the potential risks associated with each case across key risk categories to assess the potential risk severity. For example, use cases that support customer journeys, such as gen-AI-enabled chatbots for customer service, may raise risks such as bias and inequitable treatment across groups (for example, by gender and race), privacy concerns from users inputting sensitive information, and inaccuracy risks from model hallucination or outdated information (Exhibit 3).

When conducting this analysis, it's important to develop a rubric to calibrate expectations of what constitutes a high versus a medium risk across categories. Otherwise, organizations may run into disagreements driven more by individual comfort on risk levels than by objective factors. To take the example of data privacy, we typically see higher-risk examples as requiring personal or sensitive information for accurate training of the model (or higher potential for users to enter personal information in interacting with the technology). Lower-risk use cases would exhibit neither of these characteristics.

Exhibit 3

Organizations that deploy generative AI use cases can create a heat map ranking the potential severity of various categories of risk.



[1]Intellectual property.
[2]Small and medium-size enterprises.

McKinsey & Company

Using this logic, developing an application that supports an adviser in providing tailored financial advice would tend to rank higher in privacy risk exposure than would an application that automates basic contract templates.

It's essential that the executive in charge of the use case leads the initial assessment of the risks associated with it (as part of the role of the product manager in an effective operating model). This fosters the appropriate awareness of potential risks and accountability for managing them when the use case is approved for ultimate development. In addition, a cross-functional group, including business heads and members of legal and compliance functions, should review and validate the risk assessments for all use cases—and use the results as input when making decisions about use case prioritization.

Consider options for managing risks at each touchpoint

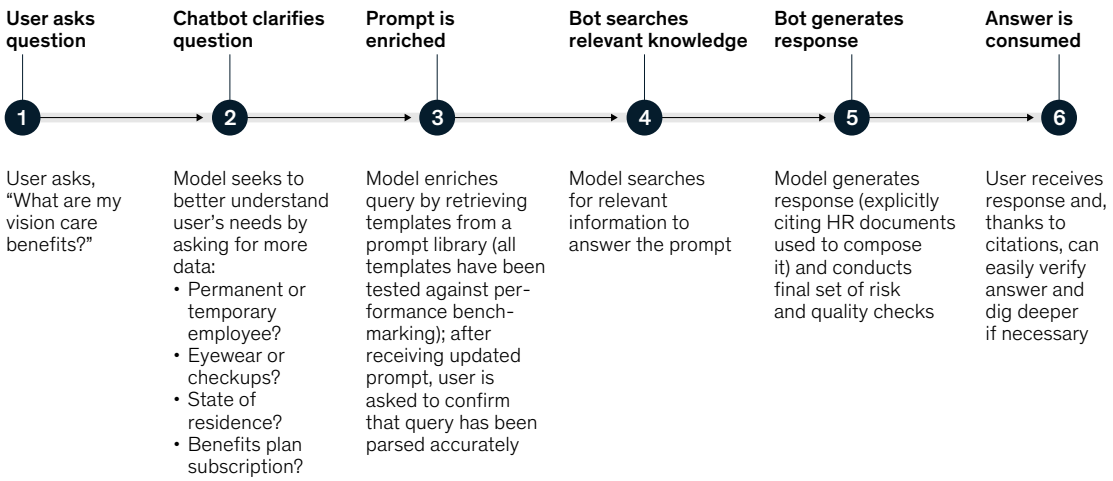
Once an organization maps the gen-AI-related risks, it must develop strategies to manage exposures through a combination of mitigation and robust governance. Many (but not all) mitigations are technical in nature and can be implemented across the life cycle of the process. Importantly, these controls don't all need to be embedded in the underlying foundation model itself (which many organizations won't have access to). Some can be overlays built in the local environment, as is the case of a gen-AI-enabled chatbot designed by an HR department to field employee queries about benefits (Exhibit 4).

In that use case, across the life cycle of a query, once a user asks a question, many possible mitigations can occur. They include having the chatbot ask clarifying questions to generate additional necessary user inputs, having the user confirm that the chatbot has properly understood the query, limiting the types of data sets that the chatbot can access (for example, excluding personal information), and designing the chatbot to provide citations to explain its answers and allow for fact-checking of its responses. Organizations implementing this use case can take steps (such as limiting repeated interactions) to frustrate the attack vectors and jailbreaking that are known to create challenges for chatbots. They can also develop classifiers to identify and reject out-of-scope queries (such as requesting calculations).

Exhibit 4

Generative AI risk can be mitigated at multiple points across a user interaction.

Sample HR chatbot interaction with built-in checkpoints to catch potential misfires



McKinsey & Company

There are important categories of additional non-technical mitigations that organizations should consider when developing use cases. At this stage of gen AI maturity, most organizations are maintaining humans in the loop to guard against the technology being able to put outputs directly into production or to engage directly with end customers. As previously referenced, contractual provisions to guard against problematic use of data from third parties are important. As a third example, organizations should develop coding standards and libraries to capture appropriate metadata and methodological standards to support reviews.

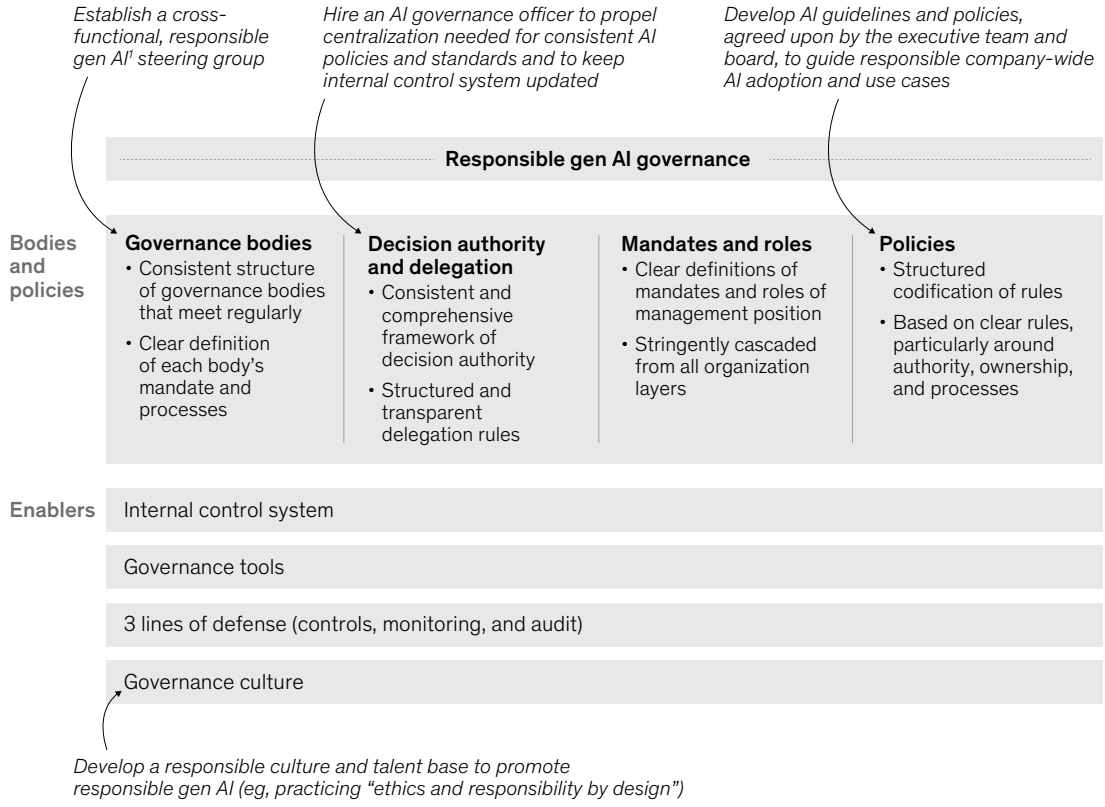
Many of the initial mitigating strategies for gen AI span multiple use cases, allowing organizations to get scaled benefits from their technical mitigations rather than having to create bespoke approaches for each case. For example, in the HR chatbot example, the ability to produce sources as part of the query answer could also be applied in use cases of an employee trying to explain a product to a customer or building analyses of peer companies. In both cases, this will go some way to addressing challenges of "explainability" and overall confidence in output.

Balancing speed to scale with judicious risk management through governance

Using gen AI will place new demands on most organizations to adapt governance structures to respond to demands on approvals and exercise oversight. However, most organizations should be able to adapt what they have today by expanding mandates or coverage (Exhibit 5). This will limit the

Exhibit 5

Moving with speed while mitigating risk often requires revised governance.



[1]Generative AI.

McKinsey & Company

potential disruption of establishing an entirely new phalanx of committees and approval bodies that could add friction to decision making and confusion over accountability.

Gen AI will likely require organizations to make changes to three core elements of governance:

— *A cross-functional, responsible gen AI steering group with at least a monthly cadence.* This group should include business and technology leaders, as well as data, privacy, legal, and compliance members. It should have a mandate for making critical decisions on managing gen AI risks, covering assessment of exposures and mitigating strategies for both inbound and adoption-based risks. It should review foundational strategy decisions, such as the selection of foundational models and compatibility with the organization's risk posture. This steering group ideally has a single individual empowered to handle coordination and agenda setting. In industries with established regulatory expectations and a long history of risk management of model and algorithmic risk (such as financial services), this person will typically be already on staff (and may be the head of model risk). For organizations facing a sudden increase in regulatory expectations from gen AI, they may need to hire an AI governance officer or similar role to discharge these responsibilities.

— *Responsible AI guidelines and policies.* Organizations should develop a set of guiding principles agreed on by the executive team and the board that will guide AI adoption and serve as a guardrail for acceptable use cases. Principles that we've seen debated include questions on the degree to which gen AI can or should be used to drive personalized marketing or customer outreach, the use of gen AI to support employment decisions (including hiring and performance reviews), and the conditions under which gen AI outputs can be put directly into production without human review. Existing policies typically need to be refreshed to account for gen AI development and use (for example, covering misrepresentation and IP infringement).

— *Responsible AI talent and culture.* A commitment to responsible AI can't rest solely in the executive ranks. Instead, it needs to cascade throughout the organization, with accountability, capability building, and awareness tailored to the relevant degree of exposure of relevant roles to the technologies. Basic organization-wide training on responsible AI should be developed and rolled out to foment a broad understanding of the dynamics of inbound risk and how to engage with the technology safely. For example, given the potential for the models to hallucinate, users should be told, as part of their training, that they shouldn't accept an answer just because their machine has provided it (in contrast to how they may have experienced prior office productivity technologies). Those engaged in the development and scaling of use cases should have a deep understanding of ethics and "responsibility by design" to embed risk considerations early in the design and engineering processes. Talent considerations include embedding a mix of nontechnical and technical talent—and ideally, technical talent with risk expertise to support identification and design of user query workflows and controls.

Implementing responsible gen AI: It's all about governance and people

Establishing the right governance is a necessary but not sufficient step in driving responsible adoption of gen AI use cases at scale. As referenced in the preceding section, embedding responsibility by design into the development process is essential for judicious deployment of the technology. There are four critical roles required for successful implementation of this throughout the use cases, where the responsibilities of these roles are tied closely to their talent and expected actions in pushing forward use cases:

— *Designers.* Designers, or product managers, steer the direction of gen AI deployment by identifying new use cases with an awareness of how they fit into the organization's overall gen AI strategy and road map. They're typically drawn from within the businesses and functions for which the organization has the most

conviction that gen AI can have significant impact. The product managers should be accountable for identifying and mitigating relevant risks. They will have an important role in driving the cultural changes required to adopt gen AI, including building trust in the proposition that business value can be achieved responsibly and safely for employees and customers.

— *Engineers.* Engineers are technical experts who understand the mechanics of gen AI. They develop or customize the technology to support the gen AI use cases. Just as important, they're responsible for guiding on the technical feasibility of mitigations and ultimately coding the mitigations to limit risk, as well as developing technical-monitoring strategies.

— *Governors.* Governors make up the teams that help establish the necessary governance, processes, and capabilities to drive responsible and safe implementation practices for gen AI. These include establishing the core risk frameworks, guardrails, and principles to guide the work of designers and engineers and challenging risk evaluation and mitigation effectiveness (especially for higher-risk use cases). The AI governance officer is a prime example of this persona, although the role will need to be complemented with others, given the range of potential risks. These roles will ideally cover data risk, data privacy, cybersecurity, regulatory compliance, and technology risk. Given the nascency of gen AI, governors will often need to coordinate with engineers to launch "red team" tests of emerging use cases built on gen AI models to identify and mitigate potential challenges.

— *Users.* Users represent the end users of new gen AI tools or use cases. They will need to be trained and acculturated to the dynamics and potential risks of the technology (including their role in responsible usage). They also play a critical role in helping identify risks from gen AI use cases, as they may experience problematic outputs in their interactions with the model.

An operating model should account for how the different personas will interact at different stages of the gen AI life cycle. There will be natural variations for each organization, depending on the specific capabilities embedded in each of the personas. For example, some organizations will have more technical capabilities in designers, meaning they may have a more active delivery role. But the intent of the operating model is to show how engagement varies at each stage of deployment.

Gen AI has the potential to redefine how people work and live. While the technology is fast developing, it comes with risks that range from concerns over the completeness of the training data to the potential of generating inaccurate or malicious outputs. Business leaders need to revise their technology playbooks and drive the integration of effective risk management from the start of their engagement with gen AI. This will allow for the application of this exciting new technology in a safe and responsible way, helping companies manage known risks (including inbound risks) while building the muscles to adapt to unanticipated risks as the capabilities and use cases of the technology expand. With major potential uplift in productivity at stake, working to scale gen AI sustainably and responsibly is essential in capturing its full benefits.

Oliver Bevan is a partner in McKinsey's Chicago office; **Michael Chui** is a partner in the Bay Area office, where **Brittany Presten** is an associate partner and **Lareina Yee** is a senior partner; and **Ida Kristensen** is a senior partner in the New York office.

Designed by McKinsey Global Publishing