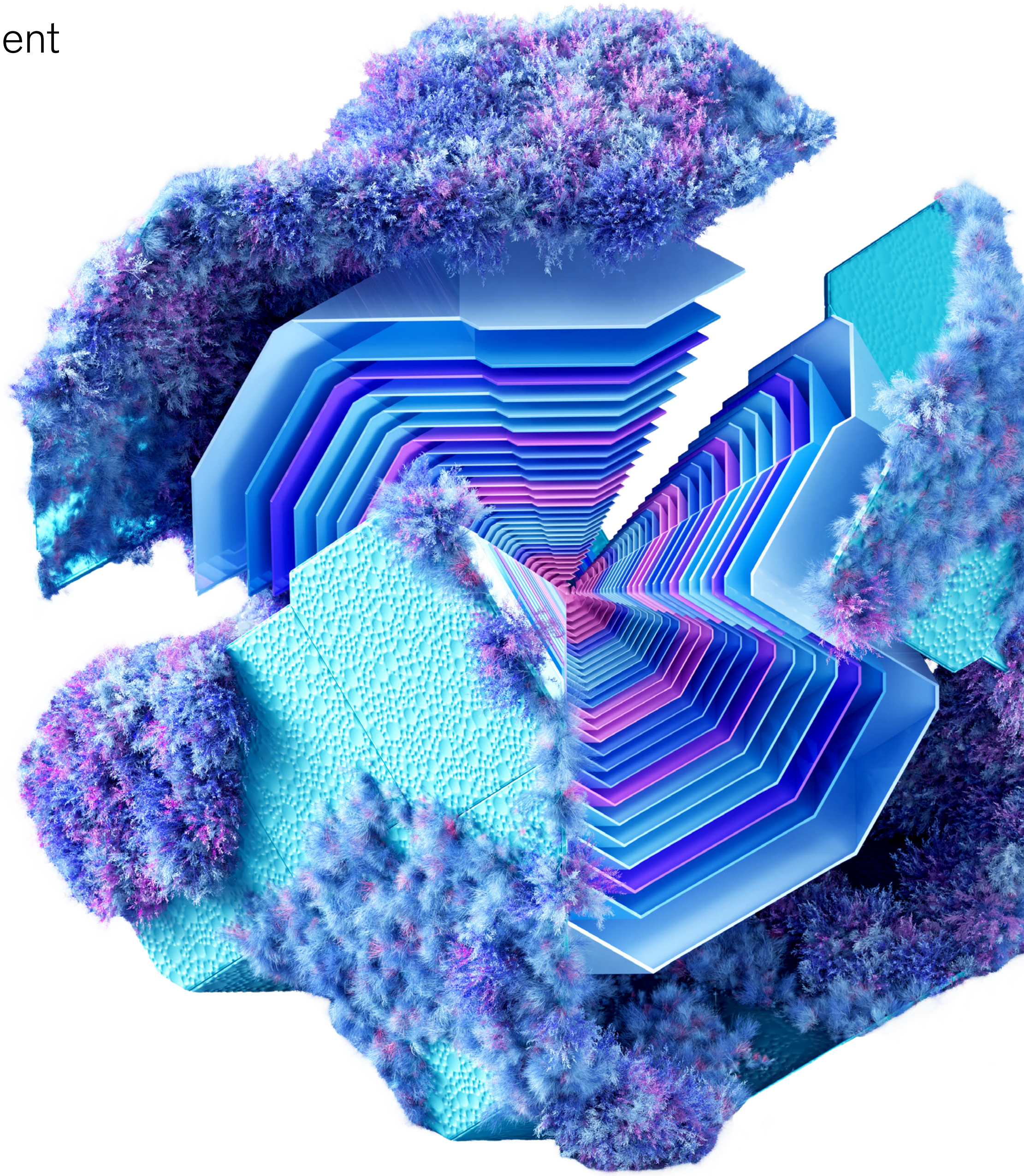
McKinsey
& Company

McKinsey on Risk & Resilience

Special edition: The new frontiers of AI in risk management



Number 20, November 2025

The articles in *McKinsey on Risk & Resilience* are written by risk experts and practitioners from McKinsey's Risk & Resilience Practice and other firm practices. This publication offers readers insights into value-creating strategies and the translation of those strategies into company performance.

This issue, and future issues, are available to registered users online at McKinsey.com. Comments and requests for copies or for permissions to republish an article can be sent via email to McKinsey_Risk@McKinsey.com.

Cover image:
© Eugene Mymrin/Getty Images

External relations,
Global Risk & Resilience Practice:
Bob Bartels

Editor: Roberta Fusaro

Contributing editor:
Joanna Pachner

Art direction and design:
LEFF

Data visualization:
Jessica Wang, Jonathon Rivait, Matt Perry, Richard Johnson

Managing editor:
Heather Byer

Editorial production:
Charmaine Rice, Dana Sand, Drew Holzfeind, Kanika Punwani, Katie Shearer, Katrina Parker, LaShon Malone, Maegan Smith, Mark Cajigao, Mary Gayen, Nancy Cohn, Pamela Norton, Pooja Yadav, Ramya D'Rozario, Regina Small, Roger Draper, Sarah Thuerk, Sneha Vats

McKinsey global publications

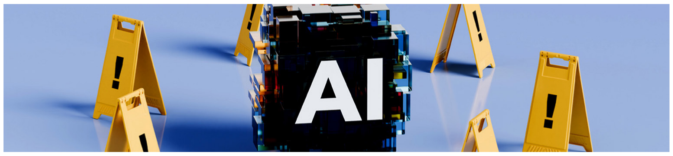
Publisher: Raju Narisetti

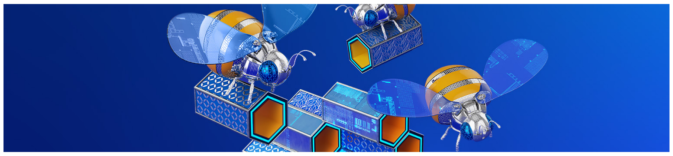
Global editorial director and deputy publisher:
Lucia Rahilly

Global Publishing board of editors: Lucia Rahilly, Mark Staples, Monica Toriello, Rick Tetzeli, Roberta Fusaro

This publication is not intended to be used as the basis for trading in the shares of any company or for undertaking any other complex or significant financial transaction without consulting appropriate professional advisers.

Contents





Introduction

Artificial intelligence, specifically agentic AI, is having a transformative impact on both financial institutions and corporate entities, bringing the role of the chief risk officer (CRO) and the risk management function front and center.

As organizations navigate complex threats and opportunities, leaders must ensure that their organizations adapt and evolve to remain resilient and competitive. Our experience has shown us that today's risk professionals are at the forefront of organizational and operational success.

In this special edition of *McKinsey on Risk & Resilience*, "The new frontiers of AI in risk management," we discuss the results of a survey about gen AI's impact on the credit business; share best practices for fighting financial crime with agentic AI; provide a comprehensive scorecard that can help companies redesign their risk governance frameworks; take a deep dive on securing agentic AI; and offer a road map to navigating the risks and opportunities of AI.

With AI, organizations understand the importance of leveraging risk management not just as a defensive measure but also as a strategic imperative. Our insights help chart a practical AI path—one that balances innovation with sound governance, transparency, and a commitment to ethical outcomes.

- *The agentic approach in AI.* The agentic approach in AI is revolutionizing productivity by using large language models (LLMs) for focused tasks and integrating them with tools and employees that interact with a company's digital ecosystem. This method surpasses the gen AI approach by enabling faster and broader digitalization, including complex, unstructured processes that are still rule-based.

- *Impact on financial institutions.* For banks and financial-services companies, the agentic approach is a game changer. It enhances productivity by digitalizing complex processes, making them more cost-effective, faster, and more reliable. This is particularly significant for the CRO, who oversees risk management and compliance.

- *The role of the CRO.* The CRO's role is evolving in response to the integration of AI. CROs must now consider a broader spectrum of model risks, facilitated by new regulations, such as the EU Act. CROs' quantitative and technological expertise positions them as natural leaders in AI development and deployment, requiring collaboration across risk, IT, and business functions.

- *Rapid development of AI tools.* The pace of AI development is accelerating, with an influx of new tools and advanced LMMs. This rapid evolution requires organizations to adapt quickly and plan strategically to stay ahead.

McKinsey, through its Risk & Resilience Practice and QuantumBlack, AI by McKinsey, is at the forefront of innovation in these areas. We collaborate with clients to navigate the complexities of AI integration, and in this issue, we share key experiences and insights as a guide. We hope you find this content useful and look forward to publishing further deep dives on AI in the near future.

Thomas Poppensieker

Thomas Poppensieker
Senior partner and chair,
Global Risk & Resilience Editorial Board

Banking on gen AI in the credit business: The route to value creation

Banks have taken steps to accelerate the adoption of gen AI in the credit business, but most remain on a long-term journey, according to a recent survey.

This article is a collaborative effort by Arvind Govindarajan, Filippo Maggi, and Kevin Buehler, with Jania Kesarwani and Maria Acuna, representing views from McKinsey's Risk & Resilience Practice.

© Getty Images

Transformative technologies don't come along very often, so when they do it pays to act quickly. When gen AI algorithms were launched in 2022, banks wasted little time exploring their potential in core commercial credit activities. But three years later, the results are mixed, with some institutions making good progress in putting the technology to work while others lag behind, a new study from McKinsey and the International Association of Credit Portfolio Managers (IACPM) shows (see sidebar, "Our methodology").

Gen AI is now a priority for many banks

To gauge banks' progress in adopting gen AI in the credit business, we interviewed and surveyed senior executives at 44 financial institutions globally. Across banks ranging in size from megaplayers to regionals, we asked about the factors affecting their adoption of gen AI, their most promising use cases, and their approaches to managing risks associated with the technology.

The responses were unequivocal on one point: Gen AI is starting to break through, with about half of senior leaders identifying it as a priority. Indeed, in key applications such as credit decisioning and pricing, rising numbers of institutions are rolling out one or more use cases. Moreover, credit applications often rank on a par or ahead of other applications, with executives seeing particular potential for gen AI in early-warning systems, credit memo drafting, and customer engagement activities.

That said, sentiment is not universally positive. Many banks are cautious about scaling amid continuing skepticism over the technology's financial benefits. As a result, only a few, mainly larger institutions are ahead of the curve, while most say progress has been slower than expected.

Survey respondents tell us there are several reasons for the industry's incrementalist approach. Many banks, for example, are still missing the skills, frameworks, and operational architectures they need to implement gen AI successfully. Underlying these challenges, we see two structural constraints: First, decision-makers are focused too narrowly on simple use cases rather than seeking to transform more complex workflows and end-to-end journeys. Second, we find that most banks have only recently started to deploy agentic AI, a version of the technology that uses decisioning algorithms to create cross-cutting impacts, for example, in the middle and front offices across lines of business. Banks that address these underlying challenges are creating competitive impetus ahead of their peers.

Our methodology

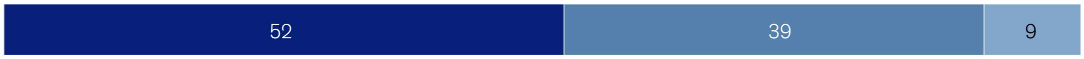
For the purposes of this article, McKinsey surveyed and interviewed decision-makers at 44 institutions globally in the second half of 2024. Our respondents included a roughly equal number of executives across megabanks, super-regionals, and core regionals. Megabanks comprised institutions with more than $1,000 billion in assets, super-regionals included institutions with $500 billion to $1,000 billion in assets, and core regionals were defined as having $100 billion to $500 billion in assets. We also connected with insurance companies/brokers and development banks.

Most institutions are testing credit use cases
Given a wide range of value creation opportunities, 52 percent of institutions have positioned gen AI adoption as a priority, our survey shows (Exhibit 1). That means senior leadership has prioritized developing gen AI use cases and backed that ambition through investment and hiring. Another 39 percent of institutions say they are interested in gen AI, but adoption is not yet a clear priority, and 9 percent admit that senior leaders are not actively engaged on the topic.

Exhibit 1

Leadership at a majority of institutions positions gen AI as a priority.

Leadership commitment to the adoption of gen AI,[1] % of respondents



Adoption is a priority

Senior leadership promotes developing gen AI use cases as a priority and supports through investments and hiring and demonstrates through tone and actions that there will be setbacks given the technology is nascent

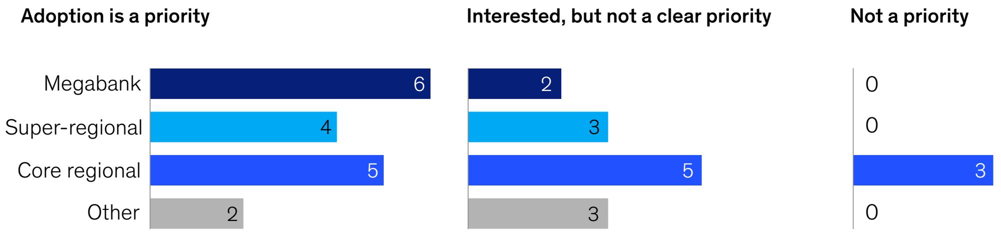
Interested, but not a clear priority

The organization is encouraged to learn about gen AI and is supportive of use case proofs of concept; however, there is less commitment to investments or hiring without a "proven" ROI and knowledge of potential setbacks

Not a priority

Senior leadership does not seem to proactively engage with the topic; the message is rather to approach with caution based on the associated risks

Commitment to implementation of gen AI, by type of institution,[1] number



[1] Question: How would you describe your institution's leadership commitment to the adoption of gen AI? (select one).
Source: IACPM and McKinsey study on the use of generative AI in credit portfolio management

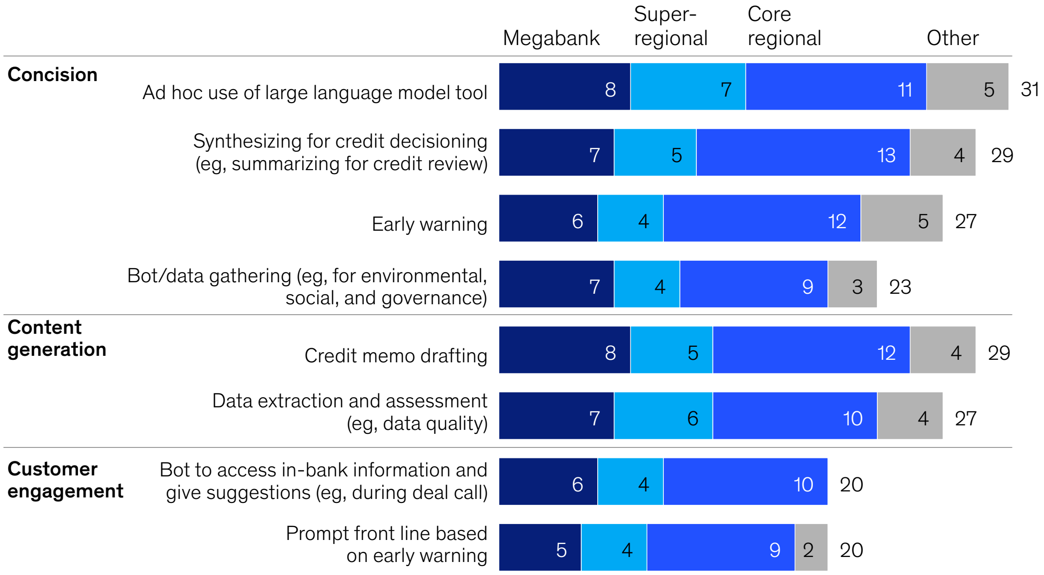
McKinsey & Company

Gen AI offers financial institutions three highly useful capabilities: concision, meaning the ability to summarize large volumes of data into digestible nuggets; content generation; and customer engagement, mainly seen in the use of bots to support relationship managers and others. Of the three, the largest number of institutions in our survey have made the most advances in concision, with the majority of institutions trying out gen AI applications in activities such as early-warning systems and credit decisioning (Exhibit 2). In one example, a multilateral development bank is exploring a gen AI tool to find the right credit-assessment documents, read and synthesize them, and draw conclusions.

Exhibit 2

Gen AI use cases in commercial credit vary by the size of the institution.

Factors for prioritizing gen AI use cases,[1] % of respondents



[1] Question: Which gen AI use cases are your institution currently implementing in commercial credit and what are their development stages? (multiple choice).
Source: IACPM and McKinsey study on the use of generative AI in credit portfolio management

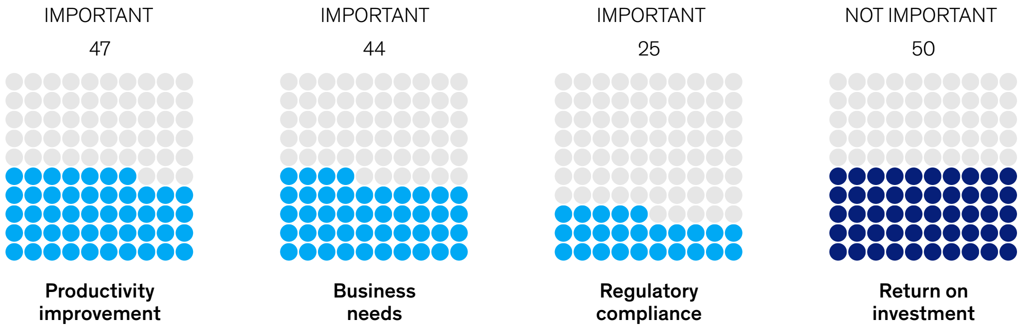
McKinsey & Company

When initiating or developing use cases, 47 percent of institutions say the most important factor is the promise of uplifts in productivity, followed closely by business needs and regulatory compliance, cited by 44 percent and 25 percent of respondents, respectively (Exhibit 3). Notably, half of institutions do not see return on investment as a major consideration, ranking it as the least important factor in making prioritization decisions. One reason may be that there are no easy ways early in the process to quantify financial impacts.

Exhibit 3

Productivity improvement is the most important factor when initiating or developing use cases.

Prioritization and importance in the initiation of gen AI use cases,[1] % of respondents



[1] Question: How would you rank the following factors in terms of their prioritization/importance in the initiation/development of gen AI use cases in your institution? (rank order).
Source: IACPM and McKinsey study on the use of generative AI in credit portfolio management

McKinsey & Company

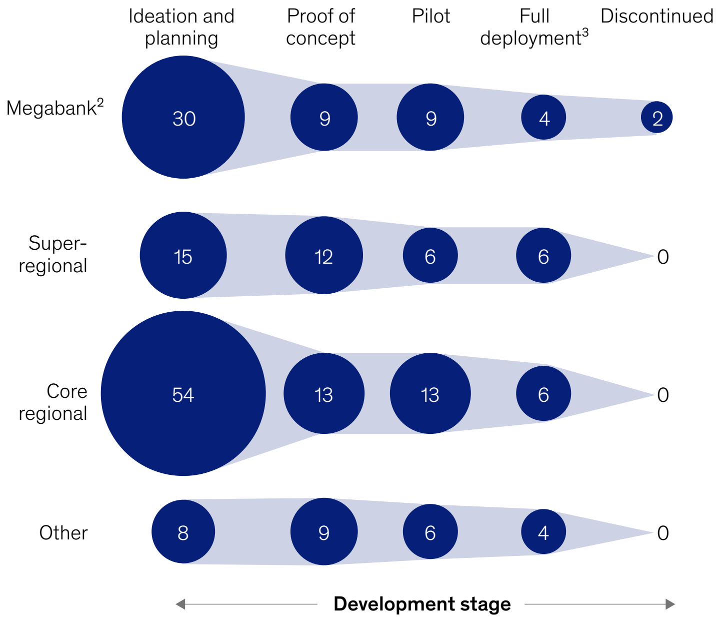
Somewhat surprisingly, the group most advanced in deployment is regional banks, which are ahead of megabanks in number of use cases (Exhibit 4). In addition, core regionals are most advanced on ideation and planning.

Very few use cases have reached the stage of full deployment, our survey shows. However, some are further advanced than others. For example, 24 percent of institutions have fully deployed use cases for “ad hoc” applications (Exhibit 5). In that context, several banks report having launched virtual LLM assistants to support use cases such as document processing (PDF conversion, digitizing) and quick QA. And while no bank has yet reached full deployment on synthesizing information for credit decisioning, 27 percent are at the piloting stage. Content generation use cases such as the drafting of credit memos and data assessment are also among the most piloted.

Exhibit 4

Regional banks are leading deployment.

Gen AI adoption, by development stage and size of institution,[1] number



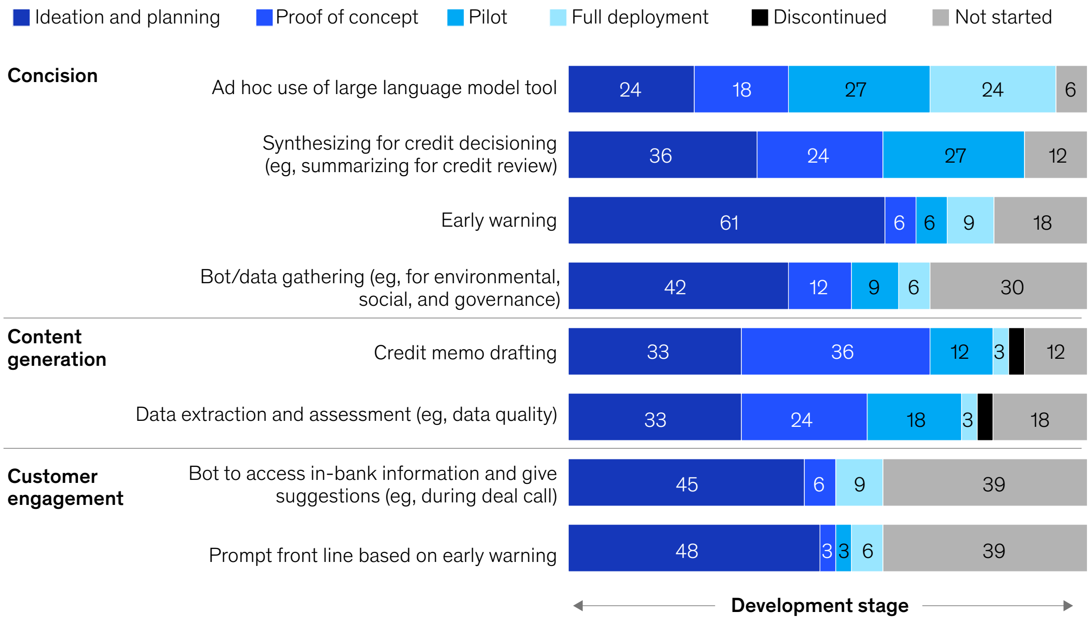
[1] Question: Which gen AI use cases are your institution currently implementing in commercial credit, and what are their development stages? (multiple choice).
[2] Megabank includes institutions with >$1,000 billion in assets; super-regional includes institutions with $500 billion to $1,000 billion in assets; core regional includes institutions with $100 billion to $500 billion in assets; other includes insurance companies/brokers and development banks.
[3] Includes optimization and maintenance and expansion and scaling.
Source: IACPM and McKinsey study on the use of generative AI in credit portfolio management

McKinsey & Company

Exhibit 5

Full deployment is rare across use cases.

Gen AI use cases in commercial credit and their development stage,[1] %



[1] Question: Which gen AI use cases are your institution currently implementing in commercial credit and what are their development stages? (multiple choice).
Source: IACPM and McKinsey study on the use of generative AI in credit portfolio management

McKinsey & Company

Why banks are taking a conservative approach
Many senior bankers, especially at regionals, are convinced that gen AI applications can create efficiencies, but there is a common gap between attitudes and implementation. Indeed, just 12 percent of North American survey respondents have deployed any use case at all.

At a McKinsey-hosted chief risk officer roundtable in 2023, we asked decision-makers what was holding them back on gen AI adoption. Sixty-seven percent highlighted shortages of gen AI capabilities, while 50 percent pointed to difficulties including defining uses cases and value at stake. A related point was that institutions putting an emphasis on early ROI from the technology were in fact more likely to give up on it, while others that pushed on through had started to see success.

Over the interim period, not too much has changed. Caution is still widespread, reflecting concern over risks that include data security breaches, model hallucinations (faulty outputs), cost-related risks, lack of validation, model and data bias, and latency issues. More than two in five institutions say they have slowed use case development because of disappointing outcomes. Reasons include insufficient accuracy and a lack of articulation on benefits. Indeed, where business scenarios require close to 100 percent accuracy, hallucinations are seen as a significant issue, while some leaders are concerned about the amount of work required to marshal data.

Forty one percent of survey respondents say that model validation issues are holding them back; one reason cited for this is the lack of historical data to assess model performance. Other constraints include too many stakeholders being involved in projects and underlying challenges that include the time and budget required (for example, to create computational intensity for development and maintenance). Upstream data risk and compliance obligations are also commonly cited as headwinds. In that context, and especially where use cases produce marginal outcomes, the path of least resistance is to proceed slowly.

All told, more than a third (36 percent) of survey respondents say they recognize gen AI's long-term potential but believe in incremental adoption. That thinking is mainly characterized by deploying smaller pilots and use cases, alongside a focus on risk mitigation ahead of scaling. Another 27 percent describe themselves as balanced but risk aware, meaning they recognize gen AI's transformative potential but remain vigilant over risks.

A final, deeper challenge goes to the fundamental issue of scope. Rather than pursuing domain-wide transformation, many banks are experimenting at a micro level and focusing on isolated use cases. In short, they underestimate gen AI's potential to reshape operations, customer engagement, and risk management.

Tackling challenges and building capabilities

Where banks are making progress is in laying the foundations for deeper gen AI adoption. Our survey shows, for example, that most institutions are in the process of attracting talent (87 percent of institutions said they are hiring technology experts, while 60 percent are training leadership teams on gen AI and its applications) and establishing secure environments and processes.

Many banks are building centers of excellence, which are tasked with developing and maintaining the architecture for gen AI applications, managing platform and deployment processes, and creating frameworks, playbooks, and guardrails. On infrastructure and technology, 31 survey respondents say they are developing and maintaining secure environments and sandboxes for experimentation. Others are running workshops, engaging outside experts, and putting in place protocols and governance frameworks as they balance experimentation and risk management.

On risk, many institutions are emphasizing data security, including establishing guardrails to prevent data exposure. They are setting up comprehensive training programs to educate users on prompt libraries and result validation. In parallel, they are embracing dedicated change management programs, human oversight of AI- and gen-AI-generated results, and stringent approval processes for use cases involving internal data or external outputs. Where use cases may impact clients or require regulatory compliance, many banks are erecting demanding approval barriers. Finally, to address hallucinations, they are conducting performance evaluations and back testing, as well as soliciting continuous user feedback.

Almost universally, institutions are engaging with third-party technology providers. Indeed, 80 percent say they have access to external solutions, with most putting in place guardrails to protect themselves, for example, by restricting access to a subset of colleagues or through internal guidelines and data security training.

More than a third (36 percent) of survey respondents say they recognize gen AI's long-term potential but believe in incremental adoption.

Taking action: Five steps to accelerate the journey

While many of the challenges first identified in 2023 are still relevant to banks' engagement with gen AI, there are signs at the margins that leading institutions are finding a way to balance risk and reward. Many banks are taking a twin approach, working both to establish basic foundations and prioritize actions that will drive adoption. Here we present five key steps in that process:

- *Align with stakeholders.* An early priority for leading institutions is to ensure that they are fully aligned with all relevant stakeholders. Externally, they proactively engage, while internally they are clear on the importance of gen AI adoption and back their views with investment in capabilities to build tools and infrastructure.

- *Standardize data to streamline deployment.* On data, leaders work hard to standardize and unify data resources, so that teams can access unstructured data, such as text documents, in one place. They also make efforts to support end-to-end experimentation and deployment, meaning they think carefully through the process—for example, to ensure optimal functionality and data flow from start to finish. They don't move forward until the application works as expected.

- *Install modular solution architecture.* To maximize the productivity of use case development and rollout, some gen AI trailblazers are putting in place modular solution architecture, meaning they are designing products with clearly defined and interchangeable components. Through this standardized approach they can pursue multiple use cases in parallel and create customizable connections across different layers.

- *Pick low-hanging fruit.* To get early wins and encourage buy in, leading companies focus initially on the least risky use cases. For example, they prioritize development of bots for internal use only, adopting a test-and-learn approach to ensure feasibility before scaling up.

- *Roll out agentic AI.* Finally, to harness real value, agentic AI can play a significant role, helping firms move from static applications such as memo drafting to compelling domain transformation. In domain transformation, an interactive orchestrating agent guides users on the process and refines outcomes based on their input. For example, in the underwriting journey, an AI agent can notify a relationship manager (RM) about a new application and generate a personalized email draft to engage the client within seconds. In client conversations, the agent can transcribe key takeaways in real time, surface relevant analytics or documents, and provide actionable insights. And post-conversation, the agent can generate a tailored to-do list, enabling the RM to efficiently prepare material for review with the credit team. When applied across the entire loan approval journey, there are even more impacts, enabling banks to optimize customer and employee experiences and drive efficiency and effectiveness at scale.

Banks have taken steps to accelerate adoption of gen AI in the credit business, but the results of our survey show that most remain on a journey. Indeed, at many institutions, there is considerable skepticism over the technology's potential to boost productivity, often reflecting previous experiences where tech rollouts did not achieve the expected gains. For that reason, leading banks are embracing a more strategic approach, ensuring they have put in place technology, talent, and operational building blocks to win the trust of stakeholders ahead of scaling. Many are also embracing agentic AI's decision-making capabilities and are seeing positive results, not just in individual business lines but across the organization.

Arvind Govindarajan is a partner in McKinsey's Boston office, where **Jania Kesarwani** is an associate partner; **Filippo Maggi** is a partner in the Milan office; **Kevin Buehler** is a senior partner in the New York office; and **Maria Acuna** is a consultant in the Miami office.

The International Association of Credit Portfolio Managers (IACPM) is a professional organization dedicated to advancing the field of credit portfolio management. Founded in 2001, the IACPM has 144 member institutions across 32 countries, including major commercial banks, investment banks, insurance companies, and asset managers. The organization focuses on research, education, and advocacy to improve the understanding and management of credit risk within financial institutions.

How agentic AI can change the way banks fight financial crime

Financial institutions are allocating significant resources to fighting financial crime, but they are generally making little progress. AI-based solutions may be an accelerator.

This article is a collaborative effort by Alexander Verhagen, Angela Luget, Olivia Conjeaud, and Vasiliki Stergiou, with Debanjan Banerjee, representing views from QuantumBlack, AI by McKinsey, and McKinsey's Financial Services and Risk & Resilience Practices.

© Getty Images

Banks are spending ever-larger sums of money on know-your-customer and anti-money-laundering (KYC/AML) activities. But there is little evidence they are getting a good return on their investments. In fact, according to Interpol, the financial industry detects only about 2 percent of global financial crime flows, despite increasing spending by up to 10 percent a year in some advanced markets between 2015 and 2022.[1] A potential solution lies in agentic AI[2]—an evolution of analytical AI technology that offers automation and productivity throughout the client life cycle (Exhibit 1).

Much of the cost of combating financial crime relates to inefficiencies in operating models and ways of working. Indeed, banks commonly assign up to 10 to 15 percent of their full-time equivalents to KYC/AML alone.[3] In parallel, automation rates are generally low amid fragmented data resources and unstandardized data sets. The result is that teams waste a lot of time on manual tasks while clients complain of tiresome interactions and lumpy processes.

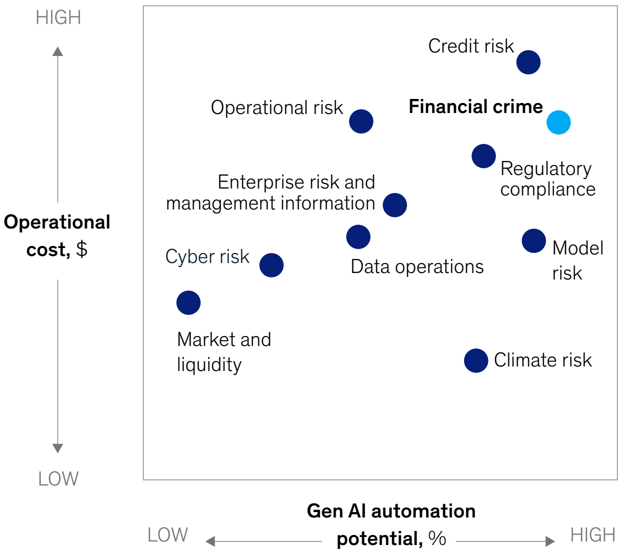
AI, specifically agentic AI, could be the antidote to KYC/AML headwinds. In this article, we map the AI landscape and examine options for implementation, highlighting how some leading institutions have deployed the technology to their advantage. Our key conclusion is that AI offers transformative potential, but only if institutions put in place the foundations and capabilities that will support an at-scale rollout.

[1] *True cost of financial crime compliance*, LexisNexis Risk Solutions, September 29, 2022.
[2] *Seizing the agentic AI advantage*, QuantumBlack, AI by McKinsey, June 13, 2025.
[3] McKinsey 2024 KYC/AML benchmark study of a set of leading North American, European, and Asian–Pacific banks.

Exhibit 1

Financial crime is a high-potential area for AI.

Operational cost and gen AI potential, by banking risk sector



Financial crime (FC) challenges

- **Large cost base:** Up to ~20% of banks' full-time employees are typically dedicated to FC activities
- **Low automation rates:** Case-handling processes lack automation and optimization, resulting in many manual reviews performed across segments
- **Data fragmentation:** Analyses depend on a mix of internal and external data, both structured and unstructured, making it difficult to deploy automated data extraction and analysis tools
- **Multitude of reports created:** FC officers spend most of their time creating detailed, case-specific reports such as know-your-client memos and negative news reports
- **Suboptimal client journeys:** Existing client processes, such as onboarding, are inefficient and fail to meet growing expectations for speed and convenience

McKinsey & Company

Analytical AI, generative AI, and agentic AI: A short tutorial on financial crime use cases

AI is not, in reality, a single technology but rather an umbrella term for a range of technologies that can understand and generate language, recognize images or speech, make decisions or predictions, and learn from data over time. In the KYC/AML context, these capabilities are broadly expressed in three forms (Exhibit 2).

Analytical AI

Analytical AI can complete analytical tasks faster and more efficiently than humans can. Prominent use cases include false positive detection in controls, including transaction monitoring, sanctions detection, name screening, and fraud detection. The technology can also produce more dynamic and integrated customer risk rating models, for example, by incorporating a higher number of behavioral (including transaction-based) factors. In transaction monitoring, it can sharpen accuracy and facilitate peer group comparisons and anomaly detection. And it can apply decision-tree-based models, a type of machine learning algorithm, to improve underperforming rules.

Generative AI

Generative AI (gen AI) learns from patterns in data sets and uses those learnings to generate original output. In KYC/AML, it can support human investigators across a number of use cases, including onboarding and in-life client reviews, based on analysis of structured and unstructured data. The technology can save human time in collecting and extracting data from documents, summarizing large sets of information (for example, on adverse media) about individuals and entities, and accelerating investigations, including analyzing purpose and nature statements, source of funds or wealth drafts, and corporate business activity descriptions. In transaction monitoring, gen AI is useful in producing alert conclusions and transaction analysis insights, supporting drafting of suspicious activity reports, and contributing to quality control and quality assurance (QA).

Exhibit 2

Three successive generations of AI development show a clear evolution in task handling.

Three examples of agentic AI use in investigating financial crimes

Traditional AI

Typically used to solve analytical tasks faster and more efficiently than humans (eg, classify, evaluate, predict, or optimize using data)

- Forecasting sales
- Segmenting customers
- Sentiment analysis

Gen AI

Used to create new content (eg, generating audio, code, images, text, and videos) and can use unstructured data more readily

- Designing concepts
- Creating marketing copy
- Generating code

Agentic AI

Has the ability to plan and execute actions against defined objectives
Can use tools (eg, web search, systems, and other models or agents)

- Context awareness and memory
- Language and communication
- Collaboration and coordination
- Tool usage
- Learning, error correction, and adaptability
- Reasoning and problem-solving
- Creativity and innovation

McKinsey & Company

In one example of the technology at work, a universal bank developed a gen-AI-driven data extraction capability to support its KYC process. The capability was deployed to production and tested with more than 50 analysts during a four-week pilot. As part of the exercise, the bank developed a reusable gen AI architecture and a codified information extraction process for more than 50 policy questions and 300 underlying subtasks. It learned that a process-first approach, based on understanding analysts' day-to-day work and involving the front line in design and testing, was an excellent way to operate.

In another use case, a large bank used gen AI to streamline generation of purpose and nature statements, as well as boost statement quality in line with bank guidelines. AI processed outputs from both raw customer data and manually created documents, significantly reducing handling times.

Agentic AI

Agentic AI refers to a technology that enables single or multiple agents to carry out tasks and make decisions autonomously (with human oversight). In the anti-financial-crime context, it is used for automating client onboarding activities, including KYC checks and refreshes, transaction monitoring, and sanctions or fraud investigations from alert to case closure.

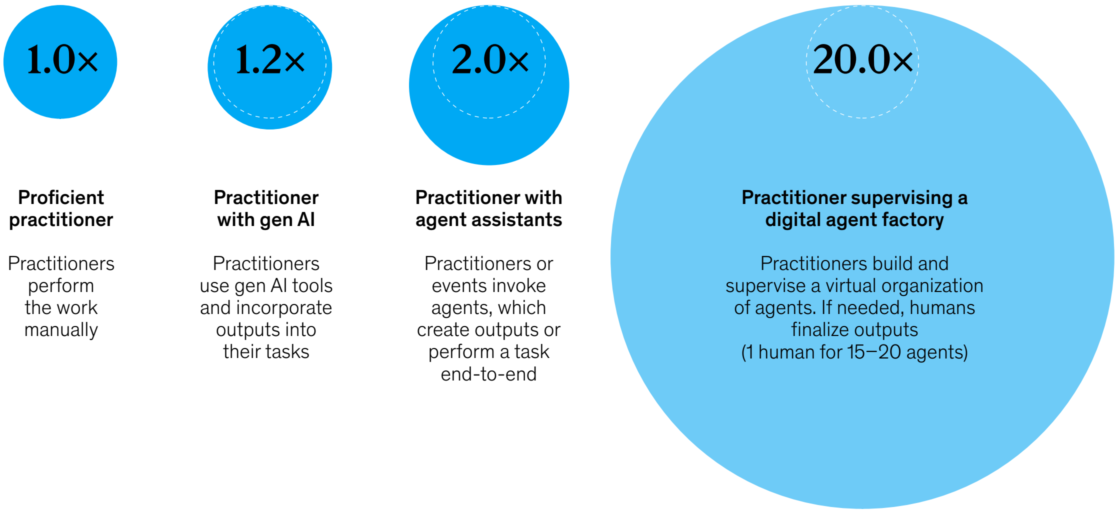
Agentic AI represents a step change in AI impact potential. While analytical AI and gen AI boost compliance efficiency and effectiveness, they often do not lead to bottom-line benefits at scale. One reason is that banks largely use them to support humans (such as KYC case handlers and transaction monitoring investigators). While this frees up time and accelerates work such as investigation handling (creating 15 to 20 percent productivity uplifts), it does not fundamentally transform effectiveness and efficiency, our experience shows.

Agentic AI, by contrast, represents a paradigm shift, with banks employing a "workforce" of AI agents (or digital factories) that can collaborate to perform end-to-end tasks autonomously. In this context, humans are only required for exception handling, oversight, and coaching (Exhibit 3). Given that each human practitioner can typically "supervise" 20 or more AI agent workers, the productivity gain can be significant—

Exhibit 3

Agentic AI offers a 20-fold increase in productivity potential.

Raw productivity potential, multiples



McKinsey & Company

anywhere from 200 to 2,000 percent, our experience shows. Banks also see a substantial positive impact on the quality and consistency of output (see sidebar, “Case study: A global bank built an agentic AI factory”).

Agents or groups of agents (squads) can be applied to different but similar tasks, for example, obtaining information on market trends and customer screening for adverse media. Here are some examples of squads employed by leading institutions in the financial crime space:

— *RAG agents* retrieve information from knowledge bases, vector databases, or document collections to answer queries with contextual accuracy. They handle embedding, chunking, and semantic search to provide grounded responses rather than hallucinated content. The agents can be used to read profit-and-loss statements, balance sheets, and company documentation to identify ultimate beneficial owners and key controllers.

— *Data pipeline agents* monitor, orchestrate, and troubleshoot extract, transform, load (ETL) processes, conduct data quality checks, and identify pipeline failures. They can automatically retry failed tasks, issue anomaly alerts, optimize resource allocation, and perform entity resolution based on analysis of customer data from different sources.

— *Research and analysis* agents gather information from multiple sources, synthesize findings, generate reports, and track emerging trends. They can monitor competitors, market conditions, or technical developments, including analyzing transactions, counterparty patterns, and alert histories.

— *Critic or validation agents* review workflow outputs, suggest improvements based on human-in-the-loop instructions, and ensure quality through to completion. They are capable of performing “self-heal and rerun” in case of minor issues (for example, input format errors).

To operate effectively, squads should always be equipped with clear boundaries, defined handoff protocols, shared content management systems, and internal guardrails.

Several principles can help banks lay the foundations

Our experience working with banks to build out AI-supported KYC/AML capabilities suggests that several principles hold true irrespective of starting position. Here are a few of the most compelling:

— *Rewire the entire domain*, including customer journeys from end to end (as opposed to individual use cases that automate individual steps within a journey).

— *Consider all levers available* to boost straight-through processing. These might include process reengineering, workflow tools, rules-based automation for simple steps, analytical AI, gen AI tools, and agentic AI to orchestrate the end-to-end journey.

— *Give AI agents distinct roles* that mirror human roles along the value chain—creating a collaborative, role-based ecosystem similar to a human team.

— *Include a QA agent in each agent squad* to check that each AI agent has completed its tasks to the required standard. In the future, agentic squads may also include compliance agents, audit agents, or other agents.

— *Redesign the operating model* to focus human expert practitioners on validation. Our experience suggests that manual intervention should be reserved only for the highest complexity exceptions and escalations (typically less than 15 to 20 percent of the total), as well as for coaching the AI agent workforce.

— *Deploy QA* for the gen AI digital factory on a sample basis, enabling a more cost-efficient approach.

Case study: A global bank built an agentic AI factory

Most financial institutions start their financial crime AI journey in either the know-your-customer (KYC) or transaction-monitoring space, reflecting the importance of those controls. One global bank was looking to reduce data processing and manual hours spent onboarding new customers, as well as move from periodic reviews to an event-driven digital customer due diligence process. It set up an agentic AI factory that encompassed an end-to-end KYC workflow—from initial KYC trigger to final memo. The bank put in place an AI agent architecture with ten agent squads. Within each squad there were four or five AI agents, including a lead agent, two or three expert-practitioner agents, and a quality assurance (QA) agent.

Each agent squad was assigned to focus on a specific step in the process and then pass the information to the next squad in the chain. Thus, one squad extracted client data from sources, including websites, annual reports, and company filings, and structured it into an initial KYC file. The second squad looked up the company in the government register, validated the country of incorporation and noted the registered shareholders and directors. A third squad performed ownership structure analysis, including identification of ultimate beneficial owners. A fourth squad performed politically exposed persons and sanctions screening on key directors and identified beneficial owners. Other squads undertook purpose and nature of relationship checks, transaction analysis checks, and adverse media screening. Finally, a stand-alone squad compiled the results into a consolidated KYC file for the human supervisor to review. The file included a summary report, a recommendation, and a detailed analysis. Depending on the results of the analysis, the agents could then choose to escalate the case further if needed.

A benefit of the agentic AI approach is that it creates a full audit trail for every agent interaction, including data used, steps followed, agent conversations, rationales for conclusions, and observations by QA, compliance, and audit agents.

How to get started: Six powerful enablers

Building a digital factory of AI agents and using them effectively on an ongoing basis requires commitment, both from the C-suite and across operations. The ideas below reflect some of the thinking we have seen driving successful outcomes:

— *Put the right people in place.* Effective implementation is contingent on KYC and risk data science skills and expertise, as well as a vision of the financial crime organization of the future. This will be predicated on identifying the required resources, including a DevOps (software development and IT operations) team, financial crime team leaders and managers, and KYC or financial crime analysts, who leverage deep domain knowledge to instruct the agent workforce, including reviewing outputs, high-end decision-making, and exception handling.

— *Be clear on the process.* Leading banks benefit from a granular and streamlined view of the target KYC or financial crime handling process and potential risks such as hallucinations or toxicity. This can prevent automation of a subpar process in the risk or compliance organization. Process flows should be broken down into distinct, independent capabilities so that banks can train and optimize gen AI bots effectively.

— *Invest in technology.* Technology is a vital element in the equation, with leading banks prioritizing the following:

  - a scalable and modular structure with access to foundation models, an enterprise agentic framework and agents repository, and APIs into internal or external data sources and applications

  - a business-friendly user interface to promote collaboration between AI agents and human supervisors, with institutions retaining existing KYC or financial crime infrastructure as much as possible

  - access to under-the-hood compute infrastructure (via cloud or on-prem) to enable AI models to operate at scale and (in some cases) in real time

— *Aim high on data.* Data quality is a primary concern for many financial institutions, and AI can help them identify and remediate data quality issues quickly. For example, in the case of sanctions, AI can support entity resolution, which is vital to identify the same customer across different data sources. Moreover, some banks are building frameworks that use AI to automatically detect, assess, and enhance data across dimensions. Key components include the following:

  - a modular architectural setup with components that can be leveraged across processes (within KYC/AML but also in business lines such as credit)

  - a clear road map for moving unstructured data (onboarding forms, policy documents, registration documents) into the analytics infrastructure and a framework of tools and AI (including agents) that monitor, detect, and report data quality issues

— *Optimize risk management.* Banks should prioritize creating a dedicated risk management framework and system for ongoing risk monitoring, including but not limited to data protection, intellectual property infringement, and hallucinations.

— *Embrace change management.* A comprehensive change management approach can guide practitioners in their new roles, for example, by providing the coaching and prompting skills needed to oversee an agentic workforce. However, the process is relatively complex, and adoption typically takes about twice as long as building the technology. Thus, leading institutions take the time to put enabling pillars in place, including redesigning underlying processes, creating appropriate roles and responsibilities, adapting the organizational structure, and establishing a talent management strategy, in which employees are evaluated against adjusted targets. Other key ingredients include timely access to data, infrastructure and large language models, as well as sandboxes and (eventually) production platforms, implemented well ahead of software development gates to avoid last-mile delays. Looking forward, banks should plan carefully for future capacity needs, especially in the front office and risk function.

Data quality is a primary concern for many financial institutions, and AI can help them identify and remediate data quality issues quickly.

The experience of leading institutions suggests AI, and especially agentic AI, could be the next major innovation lever for KYC/AML. To capture benefits quickly, leading financial institutions typically start by defining a pilot perimeter—that is, a part of the customer portfolio that they can use to experiment with a digital factory. Once impact is proven, they can prepare for scaling. Our new book, *Rewired: The McKinsey Guide to Outcompeting in the Age of Digital and AI*,[4] translates the hard-won lessons McKinsey has learned[5] helping deliver these kinds of transformations at scale.

In a rapidly changing financial crime landscape, the path to impact will likely be driven by speed of adoption (fast, at-scale model learning), a tailored operating model, and continuous maintenance of the agentic AI machine. The task should not be underestimated, but leading banks have shown that successful implementation can bring significant wins, including stronger compliance, competitive impetus, and a more streamlined customer experience.

[4] Eric Lamarre, Kate Smaje, and Rodney Zemmel, *Rewired: The McKinsey Guide to Outcompeting in the Age of Digital and AI*, Wiley, 2023.
[5] Eric Lamarre, Kate Smaje, and Rodney Zemmel, “Rewired to outcompete,” *McKinsey Quarterly*, June 20, 2023.

Alexander Verhagen is a partner in McKinsey’s Brussels office; **Angela Luget** is a partner in the London office, where **Vasiliki Stergiou** is a partner and **Debanjan Banerjee** is a principal data engineer; and **Olivia Conjeaud** is a partner in the New York office.

The authors wish to thank Barbara Jepma, Björn Nilsson, Carlo Giovine, Just van der Wolf, Sebastian Schöbl, and Sina Streicher for their contributions to this article.

How financial institutions can improve their governance of gen AI

A comprehensive scorecard can help companies redesign their risk governance frameworks and practices for gen AI and harness the power of this transformative technology.

This article is a collaborative effort by Amit Garg, David Schoeman, Gabriel Morgan Asaftei, Kevin Buehler, and Liz Grennan, representing views from McKinsey Digital and McKinsey's Financial Services and Risk & Resilience Practices.

Note: This article, originally published on March 27, 2025, has been updated for this issue of McKinsey on Risk & Resilience *with new insights based on the evolving nature of agentic AI.*

© Getty Images

Gen AI is reshaping the financial-services industry, from how banks serve customers to how executives make decisions. For all the benefits the new technology offers, including workflow automation, software enhancement, and productivity gains, gen AI also poses significant risks. It can expose a financial institution to legal and reputational risks and increase its vulnerability to cyberattacks, fraud, and more.

Trying to harness the benefits of this technology while warding off the risks can feel like a tightwire act. The heightened concerns stem from how gen AI works. Traditional AI systems are built to manage tasks that are narrow in scope by using proprietary business data. By contrast, gen AI can create new content—often by using public, unstructured, and multimodal data—through a series of complex, multistep processes that can create more opportunities for misuse and error. Traditional AI-risk-governance systems aren't designed to oversee these additional layers of complexity.

Financial institutions will need to update their AI governance frameworks to account for this increased complexity and the greater points of exposure. This will mean incorporating model risk management (MRM) and new technology, data, and legal risks into their enterprise risk model. They will need to review their oversight of AI and then assess how best to manage gen-AI-specific models going forward.

In this article, we explain how financial institutions can update and continually monitor their AI governance frameworks using a gen-AI-risk scorecard and a mix of controls. In this way, they can better identify and mitigate potential risks from gen AI and other technologies long before those risks can cause substantial financial or ethical problems.

Upgrade gen AI governance

To account for gen AI and its potential effects on business, leaders will need to systematically review all risk areas touched by the technology. They should take stock of their oversight systems, gen AI models, and intellectual property (IP) and data use, plus a range of legal and ethical factors. And now, agentic AI represents a significant leap from traditional AI and gen AI, expanding the risk surface (see sidebar, "From gen AI to agentic AI: Next-level risk governance for financial institutions").

Oversight systems

In most current arrangements, a single group (such as an MRM committee) oversees all gen AI applications. This approach typically isn't a good fit for gen AI systems, because they often comprise a blend of different models and software-like components, each of which may need specialized oversight. For example, a gen-AI-powered chatbot that provides financial advice to customers may expose companies to a range of technological, legal, and data-related risks. Accordingly, financial institutions need to decide which gen AI components only require model risk scrutiny and which require a joint review with other risk cells. Close coordination across risk committees can ensure thorough oversight.

Gen AI models

Risk leaders at financial institutions will need new models to manage gen AI risk across their companies. In the past, AI models were built primarily to do one specific task at a time, such as making predictions based on structured data and sorting data based on labels. Such tools might mine past loan data, for instance, to forecast the likelihood that an applicant might default on their loan or to identify optimal loan pricing.

With new multitasking gen AI models, banks can do more than just predict and categorize. They can devise and deliver personalized service, improve customer engagement, and enhance operational efficiency in ways that they couldn't with traditional AI. For example, gen AI models can automatically create new loan term sheets based on their analysis of similar, previously executed loans. This not only reduces manual work but also can speed up the closing process and improve the borrower's experience.

However, because gen AI models are trained on both public and private data, they can produce information or responses that are factually incorrect, misleading, or even fabricated—generating, for example, inflated income totals or an imagined

From gen AI to agentic AI: Next-level risk governance for financial institutions

As financial institutions continue to adopt AI technologies, they're progressing from gen AI to agentic AI—a more advanced form of AI that can plan, execute multistep tasks, interact with external systems, and adapt with a degree of autonomy. This evolution brings both opportunities and challenges for risk governance.

Agentic AI represents a significant leap from traditional AI and gen AI systems. While AI agents perform bounded tasks, agentic AI orchestrates complex workflows involving multiple agents and tools with key capabilities, including autonomy or self-initiating actions; complexity of multi-agents and orchestrated systems; dynamic, adaptive, and evolving behavior; and strong governance safeguards such as human-in-the-loop monitoring, kill switches, and detailed telemetry.

The shift to agentic AI expands the AI risk surface across operational risk, data and privacy risk, alignment risk, and escalation risk. To manage these risks, financial institutions need to strengthen their oversight systems. Current centralized models, such as model risk management committees, are insufficient for agentic AI's independent actions. More-dynamic scenario-based oversight is required.

To govern agentic AI effectively, institutions should consider four categories of controls:

— *Business controls:* Define clear business decision boundaries for agentic systems and establish kill-switch governance for rapid intervention when needed.

— *Procedural controls:* Implement red teaming and adversarial testing to simulate misaligned agent behavior. Conduct autonomy escalation reviews before increasing levels of delegated decision-making.

— *Manual controls:* Create specialized and targeted agentic AI oversight cells within existing AI risk governance structures. Mandate periodic audits of goal alignment and output traceability.

— *Automated controls:* Use meta-agents to monitor agentic AI behavior (AI monitoring AI). Embed self-limiting mechanisms such as capped action scopes and rollback functions.

An organization's existing gen AI risk scorecard should be expanded to include additional dimensions relevant to agentic AI, such as autonomy level, task boundaries, reversibility of actions, tool criticality, and multi-agent complexity. This enhanced scorecard will help institutions calibrate their risk posture across different AI applications.

By implementing these enhanced governance practices and controls, financial institutions can harness the power of agentic AI while maintaining customer trust, regulatory compliance, and operational safety in an increasingly agent-driven financial ecosystem.

history of bankruptcy for a customer querying a gen AI application about loan qualifications. These issues can be minimized using retrieval-augmented-generation (RAG) applications that combine external and internal data to ensure accurate responses. The RAG applications can include legally reviewed language about lending rules and can enforce strict conversation guidelines to help banks manage customers' interactions with gen AI tools.

IP and data use

Gen AI tools can introduce liabilities involving inbound and outbound IP and its oversharing. For instance, a gen AI coding assistant might suggest that a bank use computing code that has licensing issues or that may inadvertently expose the bank's proprietary algorithms. Some gen AI applications operating in real time, such as ones used in customer service, require a mix of automated and human oversight to catch issues promptly.

Many financial institutions' data governance controls don't sufficiently address gen AI, which relies heavily on combining public and private data. This raises concerns about who is responsible for what data and how it's used. For example, when using gen AI coding assistants, questions and pieces of code from open integrated development environments can be included in the prompts and sent to external gen AI providers. But they might not be saved, and their influence on code recommendations could have legal implications.

Financial institutions should develop systems to track where data originates, how it's used, and whether it adheres to privacy regulations. Not linking credit decisions to their source data could result in regulatory fines, lawsuits, and even the loss of license for noncompliance. Companies need to keep records for AI-generated content, which can change based on what's entered.

Legal and ethical factors

Headlines abound about gen AI systems that have run afoul of regulations. Mostly that's because these models blur the lines between new content and existing content protected by IP laws. This creates confusion about who owns and licenses it. Additionally, when gen AI models are trained on sensitive data, such as customer information, more attention is required for privacy and compliance. These models need careful monitoring so that they don't expose confidential information or perpetuate biases.

Transparency and "explainability" (the ability to understand how an AI model works and why it makes specific decisions) are also crucial, as the outputs of gen AI systems can sometimes be difficult to trace back to their origins. Financial institutions must establish safeguards to manage these risks throughout the model life cycle to ensure compliance with changing regulations and ethical standards.

Use a scorecard to manage gen AI risk

As financial institutions systematically review customer exposure; financial impact; the complexity of gen AI models, technologies, and data; and the legal and ethical implications, they can use a risk scorecard to determine which elements of their gen AI governance require updates and how urgent the need is. Teams can use the scorecard to evaluate the risks for all gen AI use cases and applications across the company (exhibit).

The scale used (scores of 5, 3, and 1, with 1 meaning low risk) reflects the degree of customer exposure and the level of human expert oversight in the inner workings of the gen AI application. It also reflects the expected financial impact, stage of gen-AI-application development, and more. Across these categories, oversight by human experts—particularly for high-stakes applications—is still the most effective way to ensure that gen AI systems don't make critical errors.

The scorecard can also be helpful to procurement teams in financial institutions that purchase rather than build gen AI applications; they can use it to assess their potential exposure to third-party risk and their comfort with the data and modeling techniques used by sellers of gen AI applications. While some factors may not be totally transparent to buyers, procurement teams can use a mix of vendor due diligence, technical reviews of underlying models, and contractual safeguards to assign risk scores to third-party software and make more informed purchasing decisions.

Introduce a mix of controls to govern gen AI risk

Using a risk scorecard can help financial institutions prioritize gen AI use cases based on the business need and risk/return profile of each case. Scorecards can also signal when problems arise. In both cases, the scorecard must also be supported by a risk management framework, or set of controls, for managing gen AI. Each type of control—business, procedural, manual, and automated—plays a critical role in ensuring the safe and efficient use of gen AI.

Business controls: Don't block; adjust

Financial institutions will need to design a structure that oversees gen AI risk without slowing down innovation. For example, an organization could use a centralized AI oversight committee in the early stages of adopting a chatbot or other gen AI application. Later, control could shift to a subcommittee or multiple committees. The point is to build in flexibility.

Exhibit

Teams can use a scorecard to evaluate the risks for all gen AI use cases and applications across their company.

Risk for gen AI use cases and applications, score (1 = low)[1]

	1	3	5
Customer exposure	Gen AI capabilities don't relate to customers (eg, gen AI tool that processes contracts)	Gen AI capabilities indirectly exposed to external customers (eg, gen AI application used internally to generate marketing content)	Gen AI capabilities exposed to external customers (eg, public-facing gen AI application)
Financial impact	Gen AI capabilities don't directly map to financial or operational impact	Gen AI capabilities may lead to small downside risk due to poor performance of model	Gen AI capabilities may lead to large downside risk due to poor performance of model
Model complexity	Off-the-shelf foundational model used without customization	Virtual agents built using off-the-shelf foundational models	New foundational models built or open-source foundational models retrained
Technology complexity	Gen AI applications used only as models, with no IT integration	Third-party foundational model operation tools used to build and maintain gen AI applications (eg, data platform)	Custom foundational model operation tools need to be built and maintained in gen-AI-application production (eg, tool with high degree of integration with IT)
Data complexity	Quality of training data is high, well documented, and verifiable	Quality of training data is reasonably high and well documented	Quality of training data can't be validated, quality of training data is poor, or data set includes sensitive information
Ethical risk	Gen AI data and applications have been extensively validated internally and externally	Gen AI data and applications have been extensively validated internally	Gen AI data and applications may include inherent biases or generate toxic or harmful content

[1] Scale reflects degree of customer exposure and level of human expert oversight of gen AI application.

McKinsey & Company

Companies will need to decide how risks fit into their operational models (whether centralized, federated, or decentralized) to better address new challenges posed by gen AI systems. Most financial institutions start with a centralized organizational model for gen AI risk and shift toward a partially centralized or fully decentralized model as their risk management capabilities mature. To move faster, someestablish gen AI accelerators to create consistent approaches across departments.

Procedural controls: Stay nimble

For procedures such as handling credit applications, most financial institutions should update their MRM standards. The standards should reflect gen-AI-specific risks, such as how models handle changing inputs and multistep interactions. For instance, if a bank simulates a wide range of customer responses to a virtual assistant, the MRM will need to continuously adapt. Similarly, technology review processes should be streamlined to safely integrate gen AI systems into operations. All updates should include methods for monitoring how gen AI applications adapt over time to ensure that they remain accurate and compliant as they process new prompts and new data.

Manual controls: Keep an eye on the machine
Human oversight is essential for checking sensitive outputs and ensuring the ethical use of gen AI. For example, reviewers need to redact sensitive data before models process it. When it comes to the quality of gen-AI-generated responses, financial institutions should create “golden lists” of questions for testing the models.

They should also solicit lots of feedback from customers and employees. Systems can learn from these human evaluations. The feedback can inform the accuracy and appropriateness of various outputs—for instance, how a virtual assistant “speaks” to a customer should align with institutional values and goals. The outputs should be reviewed regularly and updated as needed to bolster the models’ learning capabilities.

Automated controls: Consider third-party tools
One of the benefits of technology is that it can, in some cases, manage itself. Automated tools can sanitize data at scale, flag unusual use, and start fixes in real time. For instance, many third-party applications can remove sensitive information from documents before processing. Other third-party tools can automate vulnerability testing for gen AI systems, which helps financial institutions quickly identify and address weaknesses. Gen AI models themselves can use a combination of traditional AI and newer technologies to check their own outputs—that is, models checking models—to ensure quality control at high speeds.

As gen AI becomes an even bigger part of financial institutions, risk leaders will need to rethink how they manage the related systems. They will need to move beyond traditional AI risk practices and include real-time monitoring, robust transparency, and stronger safeguards for data privacy and ethics. A comprehensive risk scorecard and a focus on four key sets of controls can help companies find the right balance between pursuing innovation and mitigating risk. More than that, taking a systematic approach to updating gen AI risk governance can help financial institutions unlock the transformative power of this new technology to improve decision-making, customer service, and operational efficiency—and do so responsibly.

Amit Garg, **David Schoeman**, and **Kevin Buehler** are senior partners in McKinsey’s New York office, where **Gabriel Morgan Asaftei** is a partner; **Liz Grennan** is a partner in the Darien office.

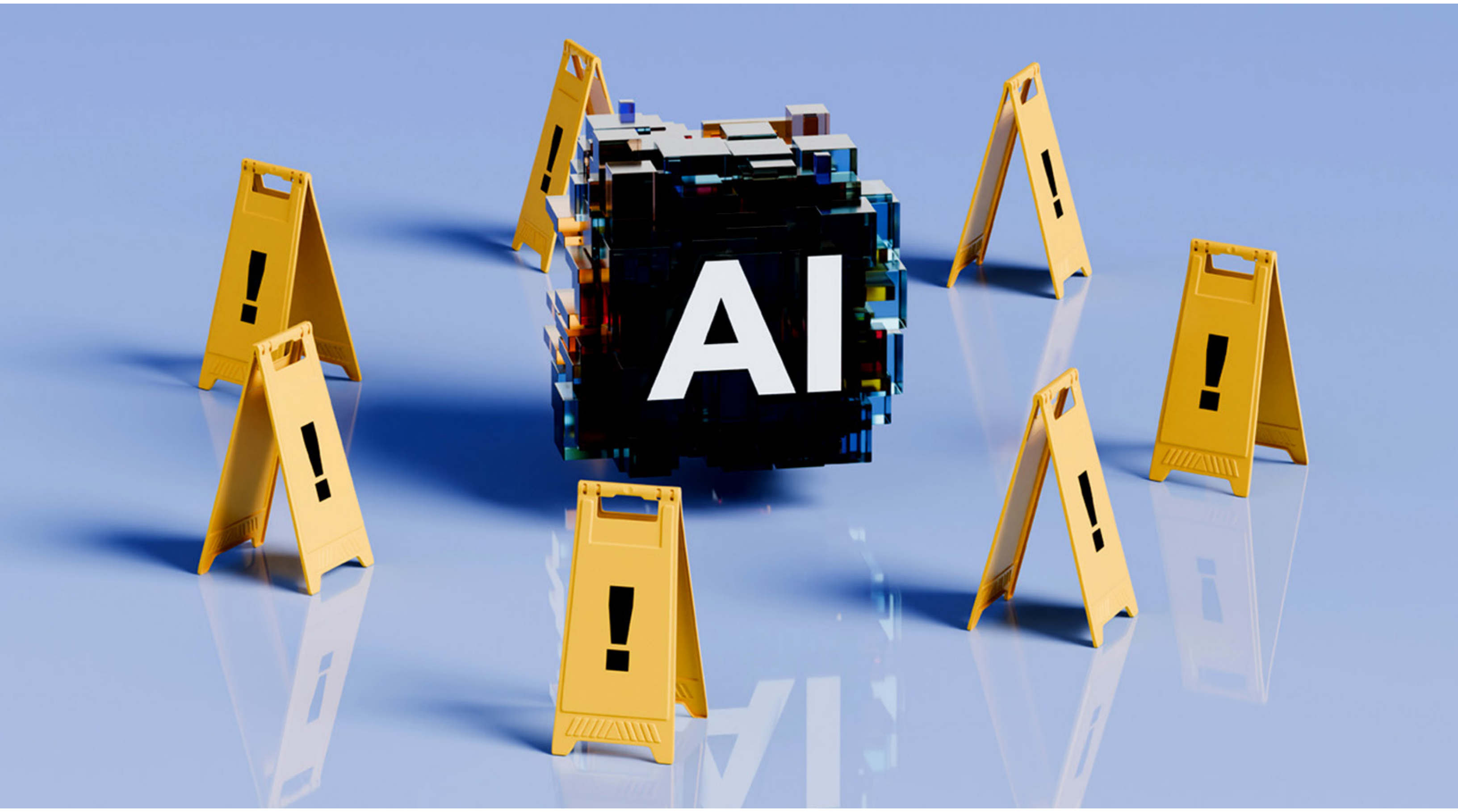
The authors wish to thank Matt Fitzpatrick for his contributions to this article.

This article was edited by David Weidner, a senior editor in the Bay Area office.

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.



© Getty Images

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]:

— *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.

[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.

- *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.

Adopting agentic AI 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.

- *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).

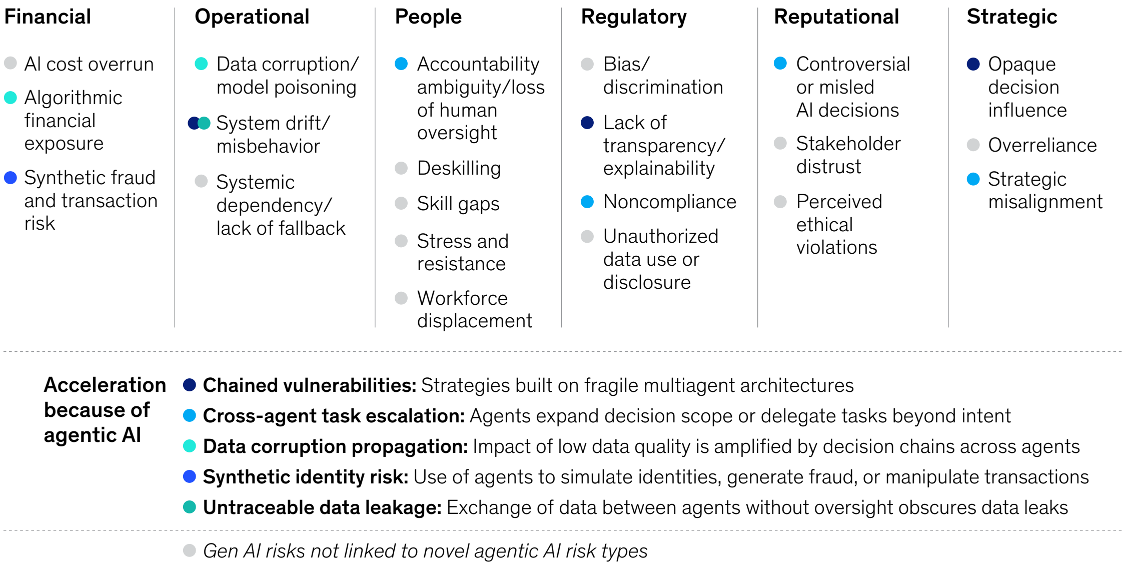
  For each agentic use case in an organization's AI portfolio, tech leaders should identify and assess the corresponding organizational risks, and, if needed, update their risk assessment methodology to be capable of measuring risks within agentic AI. Without this transparency, risks arising from agentic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

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.

Exhibit

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

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



Financial

- ● AI cost overrun
- ● Algorithmic financial exposure
- ● Synthetic fraud and transaction risk

Operational

- ● Data corruption/ model poisoning
- ●● System drift/ misbehavior
- ● Systemic dependency/ lack of fallback

People

- ● Accountability ambiguity/loss of human oversight
- ● Deskilling
- ● Skill gaps
- ● Stress and resistance
- ● Workforce displacement

Regulatory

- ● Bias/ discrimination
- ● Lack of transparency/ explainability
- ● Noncompliance
- ● Unauthorized data use or disclosure

Reputational

- ● Controversial or misled AI decisions
- ● Stakeholder distrust
- ● Perceived ethical violations

Strategic

- ● 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

— *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.

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:

— *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 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.

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

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.

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.

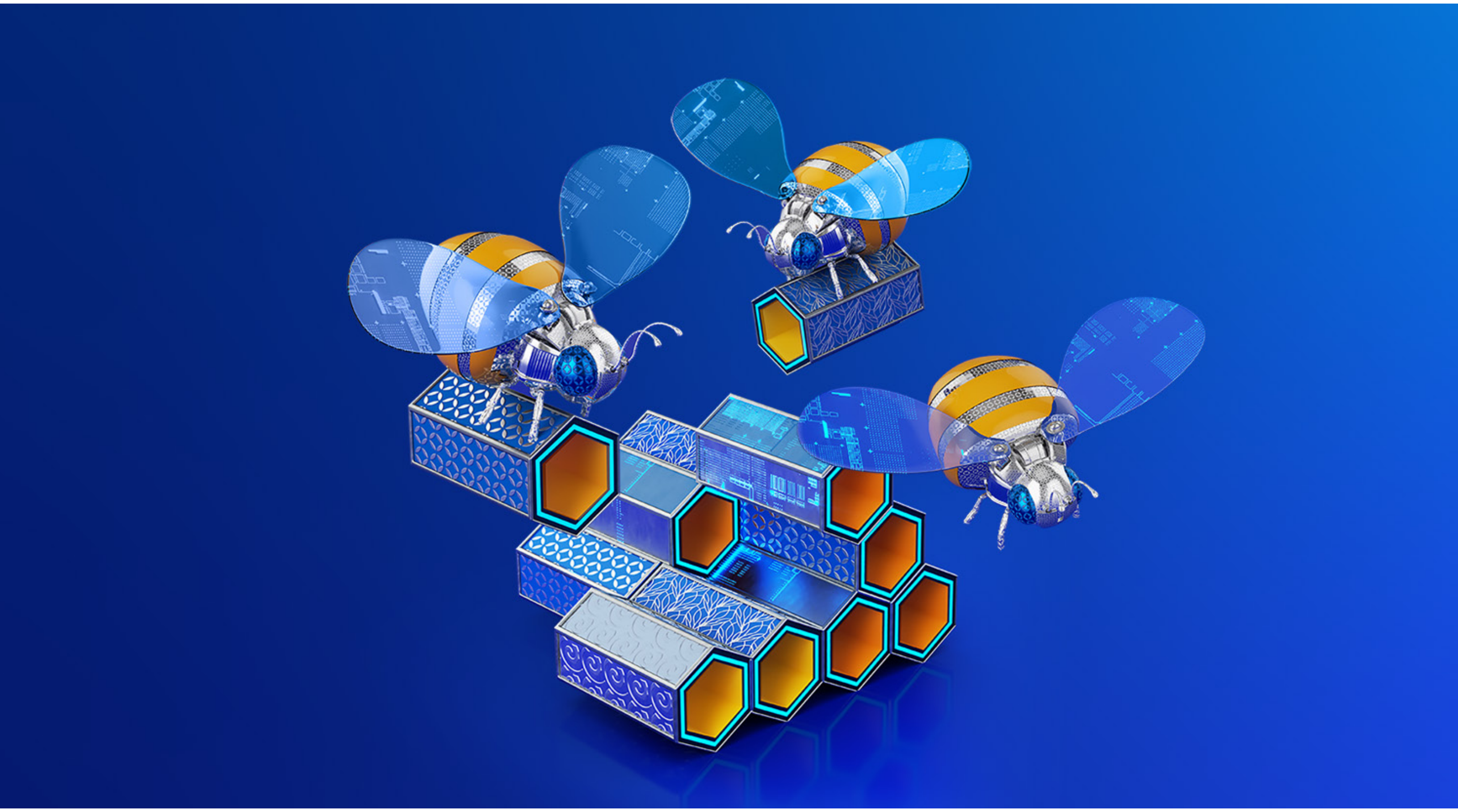
This article was edited by Larry Kanter, a senior editor in the New York office.

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.

Note: This article, originally published on March 13, 2024, has been updated for this issue of McKinsey on Risk & Resilience *with new insights based on the evolving nature of agentic AI.*



© Getty Images

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 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.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[2] Unpublished data from McKinsey survey results.
[3] *Security Intelligence*, "AI may soon defeat biometric security, even facial recognition software," blog entry by Mike Elgan, January 31, 2019.

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.

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.

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

[4] For more, see "Technology's generational moment with generative AI: A CIO and CTO guide," McKinsey, July 11, 2023.

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).

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

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

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

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. 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.

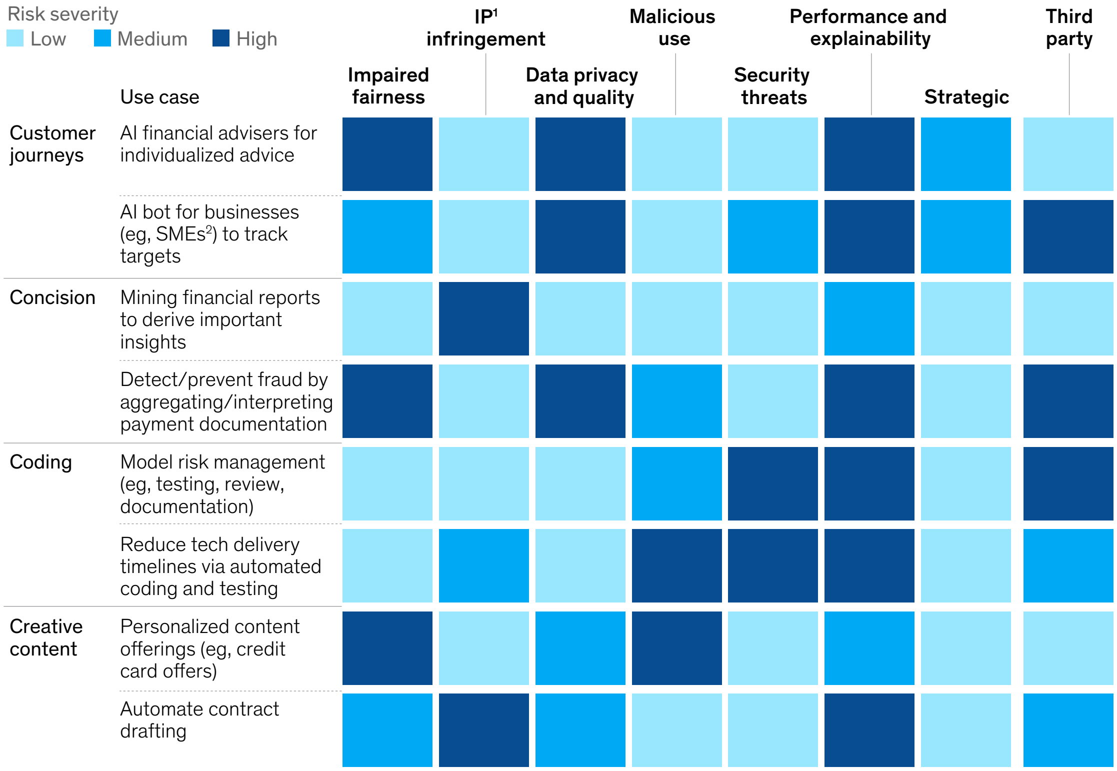
Consider options for managing risks at each touchpoint

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

Exhibit 3

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



	Use case	Impaired fairness	IP[1] infringement	Data privacy and quality	Malicious use	Security threats	Performance and explainability	Strategic	Third party
Customer journeys	AI financial advisers for individualized advice	High	Low	High	Low	Low	High	Medium	Low
	AI bot for businesses (eg, SMEs[2]) to track targets	Medium	Low	High	Low	Medium	High	Medium	High
Concision	Mining financial reports to derive important insights	Low	High	Low	Low	Low	Medium	Low	Low
	Detect/prevent fraud by aggregating/interpreting payment documentation	High	Low	High	Medium	Low	High	Low	High
Coding	Model risk management (eg, testing, review, documentation)	Low	Low	Low	Medium	High	High	Low	High
	Reduce tech delivery timelines via automated coding and testing	Low	Medium	Low	High	High	High	Low	Medium
Creative content	Personalized content offerings (eg, credit card offers)	High	Low	Medium	High	Low	Medium	Low	Low
	Automate contract drafting	Medium	High	Medium	Low	Low	High	Low	Medium

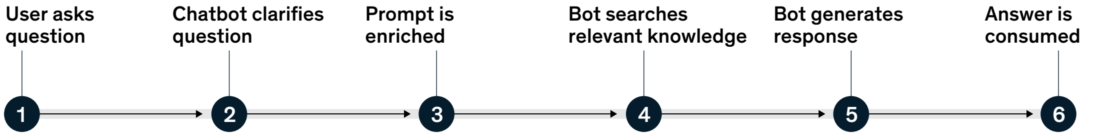
[1] Intellectual property.
[2] Small and medium-size enterprises.

McKinsey & Company

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



1. **User asks question**
User asks, "What are my vision care benefits?"

2. **Chatbot clarifies question**
Model seeks to better understand user's needs by asking for more data:
- Permanent or temporary employee?
- Eyewear or checkups?
- State of residence?
- Benefits plan subscription?

3. **Prompt is enriched**
Model enriches query by retrieving templates from a prompt library (all templates have been tested against performance benchmarking); after receiving updated prompt, user is asked to confirm that query has been parsed accurately

4. **Bot searches relevant knowledge**
Model searches for relevant information to answer the prompt

5. **Bot generates response**
Model generates response (explicitly citing HR documents used to compose it) and conducts final set of risk and quality checks

6. **Answer is consumed**
User receives response and, thanks to citations, can easily verify answer and dig deeper if necessary

McKinsey & Company

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).

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 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:

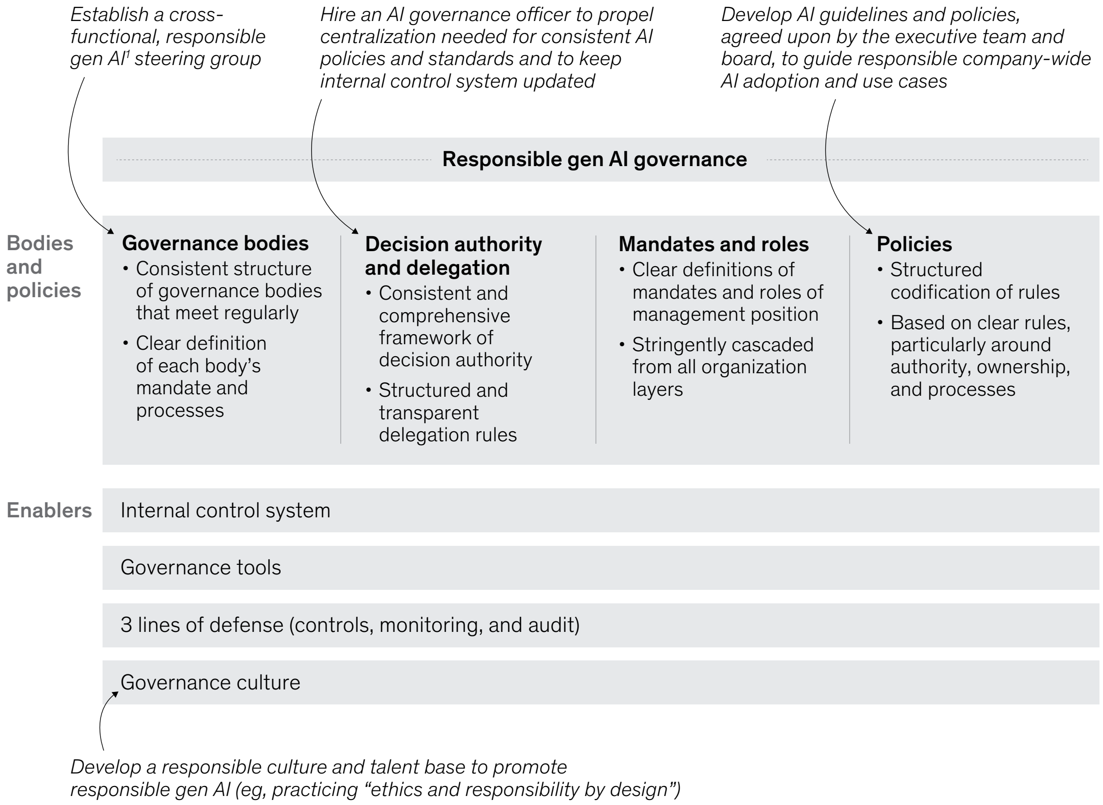
— *A cross-functional, responsible gen AI steering group with at least a monthly cadence.* This group should include business and technology leaders, as well as data, privacy, legal, and compliance members. It should have a mandate for making critical decisions on managing gen AI risks, covering assessment 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 (see sidebar "Four primary agentic AI risks"). 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

Exhibit 5

Moving with speed while mitigating risk often requires revised governance.



Establish a cross-functional, responsible gen AI[1] steering group

Hire an AI governance officer to propel centralization needed for consistent AI policies and standards and to keep internal control system updated

Develop AI guidelines and policies, agreed upon by the executive team and board, to guide responsible company-wide AI adoption and use cases

Responsible gen AI governance

Bodies and policies

Governance bodies
- Consistent structure of governance bodies that meet regularly
- Clear definition of each body's mandate and processes

Decision authority and delegation
- Consistent and comprehensive framework of decision authority
- Structured and transparent delegation rules

Mandates and roles
- Clear definitions of mandates and roles of management position
- Stringently cascaded from all organization layers

Policies
- Structured codification of rules
- Based on clear rules, particularly around authority, ownership, and processes

Enablers

Internal control system

Governance tools

3 lines of defense (controls, monitoring, and audit)

Governance culture

Develop a responsible culture and talent base to promote responsible gen AI (eg, practicing "ethics and responsibility by design")

[1] Generative AI.

McKinsey & Company

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

Four primary agentic AI risks

The increased and rapid adoption of agentic AI has led to immediate and widespread impacts on organizations. Relative to the initial stages of gen AI and reliance on chatbots, agentic AI deployment has the potential to allow for increasing workflow and task automation and joint human–AI workflows. The underlying infrastructure for creating agents is also materially simpler than what was required for creating prior AI use cases; standardized platform-based technologies allow almost anyone with access to create their own agent, practically eliminating the barrier to entry for gen AI development and usage. The implication is that for many institutions, the number of gen AI use cases will rapidly increase from dozens to thousands (if not tens of thousands) in the matter of a few months.

It has become increasingly clear that few organizations will be able to apply their initial AI risk frameworks and approaches to the agentic AI landscape without modifications. We see four primary challenges driven by the volume and complexity of agentic deployments:

— an incomplete view of how AI is being used at the institution, with corresponding challenges in aggregating trends in risk exposures and associated control strategies

— fragmented, often lengthy approach to AI governance through risk assessments, leading to difficulties standardizing risk profiles for similar agentic use cases and delayed innovation due to rigid review processes

— disconnected decision processes and unclear responsibilities, with multiple committees reviewing similar issues at different checkpoints or a lack of clarity in what is required to advance use cases to next step in review

— increasing tensions within financial institutions over whether each agent should be treated as a unique "model" and should be required to pass through model review (with associated stress on model review from the huge uptick in potential cases for review)

Based on our prior work in gen AI and recent examples of client service, our perspective is that a holistic response to these challenges requires modifying AI frameworks to increase focus on transparency and process standardization. This allows organizations to accelerate innovation through agentic deployments (and thus avoid internal frustrations with slow progress) while focusing on the risks that matter most.

To manage the increased volume and complexity of agentic AI, we are seeing our clients move to a model in which they are relaxing the need for individualized review of each use case and to a batch model that standardizes and expedites reviews based on primary usage. In practice, this means the following:

1. Organizations should establish AI use case archetypes as a common classification system for AI use, grouped by primary function. Sample archetypes we have seen are outlier detection, content generation, and data categorization. The use case owner should be accountable for identifying the primary archetype for their use case.

2. In parallel, organizations need to streamline the intake and identification mechanism for AI use cases. This typically requires understanding whether the use case constitutes a model, fits into the archetype structure, or requires a full review.

3. If the use case can be linked to an archetype, this should support the use case owner in identifying the standard set of risks and associated controls to ensure the use case development remains within acceptable risk guardrails for the organization. If the use case cannot be linked to an archetype, either because it is a novel application or it requires foundationally new infrastructure to support, it will need to go through the full risk review process.

4. Postapproval, the use case should be added to standard monitoring protocols to understand if there are potential novel risks generated or drifts from primary use cases and identify systemic implications of core infrastructure decisions (for example, if code is dependent on prior version of a model that is being replaced).

In parallel, leading organizations are increasingly linking their gen AI control framework to their underlying gen AI strategy, with an increasing focus on controls that sit at the infrastructure layer—for example, segregation of data sets that can be accessed by certain archetypes of agents but not others, access restrictions to agent archetype by employee organizational unit, and break-the-glass or kill switches for agents in case of poor performance. For established archetypes, this allows organizations to apply a standardized control set at scale without shouldering an intense burden of either compute or control design for each individual instance of AI usage.

We are starting to see increasing prevalence of agentic deployments in the risk function that will facilitate faster review and performance monitoring to identify and triage unexpected agentic behavior—much as we have sophisticated employee surveillance of higher-risk roles today. This will likely resolve some of the tensions identified above, but we expect organizations to retain a robust challenge role for the risk function that has a clear view of where higher-risk pockets of deployments are most likely to occur and how best to anticipate and resolve these.

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.

McKinsey Risk & Resilience Practice

Global coleader and North America
Ida Kristensen
Ida_Kristensen@McKinsey.com

Global coleader and Europe
Cristina Catania
Cristina_Catania@McKinsey.com

Asia–Pacific
Akash Lal
Akash_Lal@McKinsey.com

Eastern Europe, Middle East, and North Africa
Elias Hajj
Elias_Hajj@McKinsey.com

Theodore Pepanides
Theodore_Pepanides@McKinsey.com

Latin America
Cristian Berner
Cristian_Berner@McKinsey.com

Chair, Risk & Resilience Editorial Board
Thomas Poppensieker
Thomas_Poppensieker@McKinsey.com

Leader, Risk Knowledge
Lorenzo Serino
Lorenzo_Serino@McKinsey.com

In this issue

November 2025
Designed by LEFF

McKinsey.com