MAPPING THE GENERATIVE AI RISK SPACE

Nick van der Meulen, *Research Scientist*
MIT Center for Information Systems Research (CISR)

Hippolyte Lefebvre, Assistant Professor in Management Information Systems, University College Dublin *and Research Collaborator, MIT CISR*

Barbara H. Wixom, *Principal Research Scientist*
MIT Center for Information Systems Research (CISR)

The risks entailed in using generative artificial intelligence (GenAI) have increased markedly as organizations have progressed from experimentation with GenAI tools, to deployment of GenAI solutions,[1] and finally to exploration of AI agents.[2] Exposure to these risks has been amplified by massive, decentralized demand. GenAI's natural language interface makes it easy to use; its cloud infrastructure makes it ubiquitously available at low cost; and its general-purpose nature makes it applicable to myriad tasks.

To govern GenAI effectively, leaders must first understand where the risks emerge. This briefing therefore explores the **Generative AI risk space**: the set of components that give rise to risk, from the training data that shapes model behavior to the human decisions that determine how model outputs are used (see figure 1). Drawing on sixty-two semi-structured interviews with data and technology executives,[3] we describe these risks and offer guidance for leaders seeking to realize GenAI's value safely.

IDENTIFY RISKS IN CORE GENAI COMPONENTS

To make GenAI risks concrete, consider a specific example: a hiring manager using GenAI to draft a job description for a new role. What seems like a simple task—type a request, receive a polished draft—involves multiple components, each introducing distinct risks.

Training data: Foundation models have been trained on massive datasets, primarily scraped from across the internet. For our hiring manager, this means the model has absorbed millions of job descriptions and résumés, but also outdated HR practices, biased language, and inaccuracies. As such, the risk is that the model can confidently generate requirements that are incorrect for the hiring manager's industry or region, or that the output reflects outdated norms rather than current best practices. The organization doesn't control the training data; it inherits whatever the model's developer used.

Foundation model: The foundation model (typically a large language model, or LLM) distills training data into patterns it uses to generate responses. These models are inherently probabilistic, meaning that if the hiring manager provides the exact same input to the model multiple times they'll obtain different outputs each time. Foundation models also run the risk of generating "hallucinations": plausible-sounding content that is factually wrong. And because the reasoning inside these models is opaque, organizations can't determine *why* a particular output was generated—which makes such errors difficult to diagnose and correct.

Models also change significantly as vendors push updates, sometimes without adequate notice. One executive described the challenge of keeping pace: "[Model provider] employees are not allowed to present a deck to a customer if it is more than 48 hours old... Because [the provider is] moving so fast, the information in that deck is no longer current." The risk is thus that what works today may fail after an update.

User prompt: For GenAI tools to be effective, users must know how to properly direct ("prompt") the model. The hiring manager needs to provide the right context for the job description, as well as clear instructions and examples of what a good job description looks like. Without clear direction, the model's output likely won't meet expectations.

Users can also introduce risks unknowingly. When using public GenAI tools, users may inadvertently disclose sensitive information by including confidential data, proprietary strategies, or personally identifiable information in prompts. Or the user could fall victim to prompt injection, where malicious content hidden in documents or websites manipulates the model's behavior. For our hiring manager, copying text from

1 For more on GenAI tools and how they differ from GenAI solutions, see N. van der Meulen and B. H. Wixom, "Managing the Two Faces of Generative AI," MIT CISR Research Briefing, Vol. XXIV, No. 9, September 2024.

2 An AI agent is a bounded autonomous system that uses observations from its environment to select and execute actions in pursuit of specified goals. In contemporary business applications, this is often implemented using a large language model that orchestrates tool use and action sequences under defined constraints.

3 The authors conducted the interviews between Q1 and Q2 2025 with members of the MIT CISR Data Research Advisory Board from forty-one organizations.

© 2026 MIT Center for Information Systems Research, Van der Meulen, Lefebvre, and Wixom. MIT CISR Research Briefings are published monthly to update the center's member organizations on current research projects.

a compromised template could inject instructions that cause the model to leak information or act against their intent.

System prompt: Enterprise-grade GenAI tools and solutions typically combine user prompts with a hidden system prompt to enforce organizational context and constraints. This system prompt functions as the model's "constitution," setting tone, operational logic, and safety guardrails that supersede individual prompts. As system prompts govern every interaction, they require careful design and regular review. A poorly engineered system prompt creates a single point of failure that can lead to errors, security vulnerabilities, or hallucinations across the organization. A rigid, legally compliant system prompt could sanitize our hiring manager's directions, for instance, resulting in a generic boilerplate job description that lacks the excitement needed to attract top candidates.

System prompts are also vulnerable to prompt leakage. Through clever questioning, users—or bad actors—can coax the model into revealing its hidden instructions. Once exposed, competitors can replicate proprietary logic, and attackers can craft inputs designed to circumvent the organization's security guardrails.

Output: The model can generate a job description in seconds. It looks professional and appears authoritative—but it may contain biased language, inflated requirements, or fabricated qualifications. The polished appearance makes it harder for the hiring manager to spot errors. Yet proper evaluation of the output is essential, as the cost of errors can be significant. As one executive mentioned, "The probability of hallucination may be low, but the negative consequences of hallucination are really high for us." That is why some organizations enforce this directly: "Do you have the expertise to challenge the output? If not, you're not allowed to use GenAI."

Use: Finally, the hiring manager decides what to do with the output. They might use the output as inspiration, edit the model-drafted description, or accept it wholesale and post it as presented. Errors that remain internal can be caught and corrected; errors that reach external stakeholders can damage trust and reputation. Qualified candidates may not apply if AI-generated requirements seem unrealistic, or they may recognize AI-generated content and question the organization's judgment. The hiring manager's decision to post without careful review could cost the organization the very talent it sought to attract.

GAUGE EXPOSURE FROM RAG AND AGENTS

The hiring manager example illustrates risks present in any GenAI use. But many organizations have moved beyond off-the-shelf tools and solutions to deploy GenAI solutions that integrate with proprietary data and systems, expanding the risk space further.

Data assets and vector databases: To provide additional context to a foundation model, organizations can use retrieval-augmented generation (RAG) to dynamically retrieve relevant information from internal data assets[4] (often indexed in vector databases[5]). For example, the hiring manager's organization might enhance a GenAI solution with its existing job descriptions, compensation data, and HR policies. However, RAG inherits whatever quality issues exist in the underlying data. One executive described the hard lesson: "The assumptions we walked in with were that all our data was in one place and all the data was good. We spent inordinate amounts of time engineering our data to get it to work."

RAG can also expose access control gaps that previously went unnoticed. Vector databases may surface data that users could technically access but previously had no way to find. The hiring manager querying for salary benchmarks might inadvertently retrieve executive compensation data, not because RAG granted unauthorized access but because it exposed pre-existing gaps. As one executive explained, "Sometimes it turns out that the user has access to more than they knew. It can show you that data inadvertently, when previously it wasn't as easily discoverable."

Agents and tools: The latest evolution involves AI agents: software that takes autonomous actions within defined boundaries, such as querying databases, calling APIs, or executing multistep workflows. A job description generator agent might automatically pull requirements from similar roles, check salary data, and post to multiple job sites without human intervention. As agents access multiple tools and data sources, organizations can lose visibility into what data flows where and what decisions are being made on their behalf. "We need to make sure AI-based recommendations are not generated in a vacuum, or acted on autonomously," one executive warned. "The minute we take an action without that verification, we're on the hook."

The risks multiply when multiple agents coordinate to achieve complex objectives. A key risk is autonomy creep: the gradual expansion of what agents are authorized to do as organizations delegate tasks incrementally, without revisiting who oversees these actions or who is accountable for their consequences. As tasks flow between agents, the aggregate result can involve consequential decisions that no one explicitly authorized as a whole. Permissions compound the problem: access granted for one task may be reused across agents, enabling actions that were never evaluated in combination.

4 A data asset is a cohesive set of data that is made readily usable.

5 A vector database stores content as vectors (numerical representations of meaning) and is optimized to quickly retrieve the most related items using semantic similarity rather than exact keyword matches.

DISTINGUISH EMBEDDED FROM ENACTED RISKS

We identified two types of risk in the GenAI risk space, each requiring a different management approach.

Embedded risks are part and parcel of the GenAI technology and models an organization adopts. These risks exist before the organization makes any deployment decisions and are shaped by training data quality, model behavior, and performance drift introduced by vendor updates. They are not fully under an organization's direct control, and the fundamental opacity of how models reason makes them difficult to interpret and predict. Leaders cannot eliminate these risks, but they can manage them through careful vendor selection, independent evaluation, contractual requirements for transparency and change notification, and—critically—proactive vendor engagement. As one executive noted, "For the first time, because of AI, we have quarterly roadmap meetings with our vendors. We never used to do this anywhere near as actively."

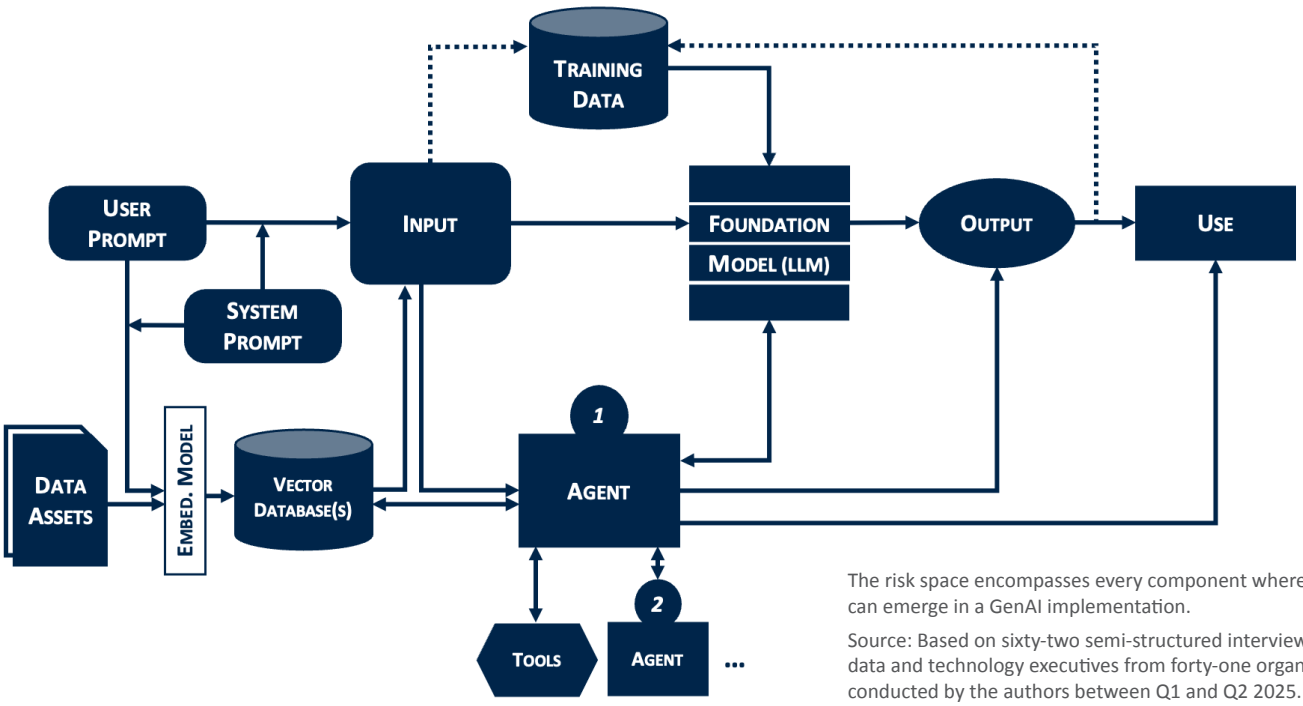
Enacted risks emerge from choices organizations make about how they deploy, configure, and use GenAI: how to design system prompts, what data users may include in prompts, what safeguards to implement against attacks, whether to require verification before acting on outputs, what permissions to grant agents, and which use contexts are appropriate. These choices reflect the quality of the organization's capabilities and governance. Managing enacted risks requires coordinated governance mechanisms—from guiding principles to mechanisms built into technical platforms—which we will explore in an upcoming companion briefing.

INVENTORY, ASSIGN, AND AUDIT

Leaders seeking to manage GenAI risks should begin by mapping their organization's exposure across the entire risk space. This means creating an inventory of every GenAI tool and solution in use. For each, document which foundation model powers it, how system prompts are designed and maintained, what data assets it connects to (and whether access controls have been verified), who may use it, what use contexts are authorized, where human review is required, who is accountable for its use, and—for agents—what permissions have been granted.

In addition, leaders should distinguish the management of embedded risks from that of enacted risks. Embedded risks require proactive engagement with vendors and partners, while enacted risks require building internal capabilities and controls. For both types, assign clear ownership of ongoing risk assessment, and establish audit trails that log the prompts, the outputs, and the human interventions at each step where GenAI influences decisions. Even if the foundation model is a black box, the GenAI risk space cannot be.

Figure 1: The Generative AI Risk Space



The risk space encompasses every component where risk can emerge in a GenAI implementation.

Source: Based on sixty-two semi-structured interviews with data and technology executives from forty-one organizations conducted by the authors between Q1 and Q2 2025.

MIT CENTER FOR INFORMATION SYSTEMS RESEARCH (CISR)

MIT CISR helps executives meet the challenge of leading increasingly digital and data-driven organizations. We provide insights on how organizations effectively realize value from approaches such as digital business transformation, data monetization, business ecosystems, and the digital workplace. Founded in 1974 and grounded in MIT's tradition of combining academic knowledge and practical purpose, we work directly with digital leaders, executives, and boards to develop our insights. Our research is funded by member organizations that support our work and participate in our consortium.

CISR RESEARCH PATRONS

AlixPartners
Avanade
Cognizant
Collibra
IFS

CISR SPONSORS

ABN Group (Australia)
Alcon Vision
ANZ Banking Group (Australia)
AustralianSuper
Banco Bradesco S.A. (Brazil)
Banco do Brasil S.A.
Barclays (UK)
BNP Paribas (France)
Bupa (Australia)
Caterpillar, Inc.
Cemex (Mexico)
Cencora
CIBC (Canada)
Cochlear Limited (Australia)
Commonwealth Superannuation Corp. (Australia)
Cuscal Limited (Australia)
Dawn Foods
DBS Bank Ltd. (Singapore)
Doosan Corporation (Korea)
Ericsson (Sweden)
Fidelity Investments
Fomento Economico Mexicano, S.A.B., de C.V.
Genentech
Hunter Water (Australia)
International Motors
Jewelers Mutual
JPMorgan Chase
Kaiser Permanente
Keurig Dr Pepper
King & Wood Mallesons (Australia)
Mater Private Hospital (Ireland)
Nasdaq, Inc.
Nomura Holdings, Inc. (Japan)
Nomura Research Institute, Ltd. Systems Consulting Division (Japan)
Novo Nordisk A/S (Denmark)
OCP Group
Pacific Life Insurance Company
Pentagon Federal Credit Union
Posten Bring AS (Norway)
Principal Life Insurance Company
Ralliant
Reserve Bank of Austrailia
RTX
Saint-Gobain
Scentre Group Limited (Australia)
Schneider Electric Industries SAS (France)
TabCorp Holdings (Australia)
Telstra Limited (Australia)
Terumo Corporation (Japan)
UniSuper Management Pty Ltd (Australia)
Uniting (Australia)
Vanguard
WestRock Company
Wolters Kluwer Financial & Corporate Compliance
Xenco Medical
Zoetis Services LLC

CISR ASSOCIATE MEMBERS

MIT CISR wishes to thank all of our associate members for their support and contributions.

Information as of January 2026

MIT CISR is funded by our members, and we gratefully acknowledge their financial support and their many contributions to our work.

Membership and benefits: cisr.mit.edu/content/member-benefits

MIT CISR research publications: cisr.mit.edu/research-library

MIT Sloan School of Management
Center for Information Systems Research

245 First Street, E94-15th Floor
Cambridge, MA 02142
t 617-253-2348 | e cisr@mit.edu

cisr.mit.edu |

Team | Isobela Byerly-Chapman, Margherita Di Pinto, Christine G. Foglia *Associate Director*, Dorothea Gray-Papastathis, Cheryl A. Miller, Ina M. Sebastian, Alan Thorogood, Nick van der Meulen, Peter Weill *Chairman*, Barbara H. Wixom, Stephanie L. Woerner *Director*