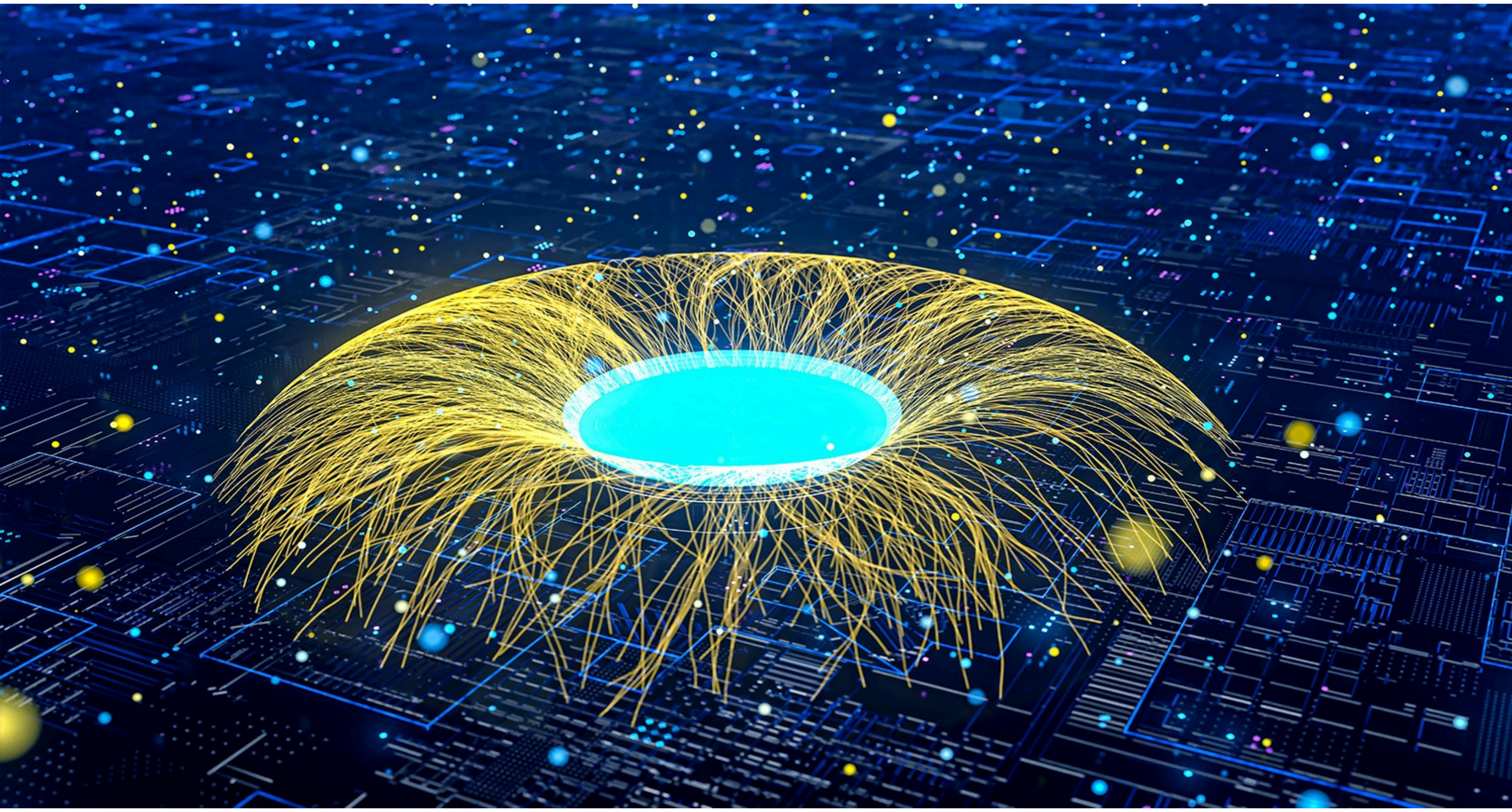
McKinsey & Company | QuantumBlack AI by McKinsey

McKinsey Technology

A data leader's operating guide to scaling gen AI

Deploying generative AI in the enterprise requires a data-centric road map. Leaders can use a well-defined operating model to successfully scale the technology.

This article is a collaborative effort by Alex Singla, Asin Tavakoli, Holger Harreis, Kayvaun Rowshankish, and Klemens Hjartar, with Gaspard Fouilland and Olivier Fournier, representing views from McKinsey Technology and QuantumBlack, AI by McKinsey.



September 2024

After almost two years of infatuation with generative AI (gen AI), companies are moving past the honeymoon phase[1] to embrace the work that matters most: creating value from this tantalizing technology. Expectations are high. A recent McKinsey Global Survey found that 65 percent of companies across sizes, geographies, and industries now use gen AI regularly, twice as many as last year.[2] Investment in gen AI continues to rise amid the belief that early gains seen by high performers are a harbinger of cost decreases and profits to come. But most companies have not yet seen significant impact from gen AI.

To keep up with the competitive pace of innovation, data executives at most organizations have drafted gen AI strategies. Not all companies have moved past the pilot stage, but most have made steps to integrate AI into their tech stacks at some level. Yet a technical integration model is only part of what is necessary to generate lasting value from gen AI. Companies must also create gen AI operating models to ensure their technology implementations deliver measurable business results.

An operating model is a familiar structure in most large organizations. A company's operating model is a plan that outlines how people, processes, and technology will be deployed to provide value to customers and stakeholders. It can encompass financial structures, partnerships, and product road maps to meet the company's long-term goals. When applied specifically to gen AI, an operating model includes every decision—from staffing and organizational structures to technology development and compliance—that guides how gen AI is used and measured throughout a company.

A well-defined gen AI operating model can help leaders successfully and securely scale gen AI across their organizations. Data is the backbone of a successful gen AI deployment, so chief data officers (CDOs) often lead the charge to create these models—bringing technology, people, and processes together to transform gen AI's potential into real impact. Yet when creating gen AI operating models, data leaders commonly fall into two traps:

— *Tech for tech:* This approach involves allocating significant resources toward gen AI without a clear business purpose, leading to solutions disconnected from real-world impact. This can result in overspending on gen AI tools that are rarely used in daily workflows and create little business value.

— *Trial and error:* This approach entails experimenting with disparate gen AI projects, but not doing so in a coordinated manner. This presents a particular risk in sectors such as technology, retail, and banking, where gen AI has the potential to quickly increase productivity. Companies in industries where gen AI may take longer to have a significant effect on productivity, such as agriculture and manufacturing, could potentially afford to wait to deploy the technology.

Many business leaders feel a sense of urgency to deploy gen AI. This creates an opportunity for data executives to get approval for gen AI operating models that put data at the center of the organization.

When CDOs and their executive supporters are ready to define a gen AI operating model, what are the first steps to get started? And what measures should companies take to ensure these operating models meet risk, governance, security, and compliance measures? We present a practical guide data leaders can use to create a gen AI operating model, including how to structure talent teams, organize data assets, and determine whether a centralized or domain-centric development is the best approach.

[1] "Moving past gen AI's honeymoon phase: Seven hard truths for CIOs to get from pilot to scale," McKinsey, May 13, 2024.
[2] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024.

About McKinsey Technology

McKinsey Technology works with organizations across the private, public, and social sectors to create opportunities around cloud, data transformation, tech strategy, risk and remediation, and artificial intelligence. McKinsey Technology comprises more than 2,500 people in every geography—developers, engineers, architects, strategists, and analysts—who bring together technical expertise and deep industry knowledge. We work closely with organizations to deliver the technology transformation they need today and build the capabilities to embrace what comes next.

Design a gen AI operating model around components

Gen AI innovation is moving at an exceedingly fast pace, so it makes sense to design an operating model that leverages components. With this approach, a company creates a plan for adding new gen AI components to the enterprise architecture at regular intervals, and in ways that are aligned with business goals. The operating model enables changes to gen AI components without having to overhaul the tech stack.

On one hand, adding gen AI functionality to mature elements that require fewer regular updates, such as cloud hosting and data chunking, warrants a higher level of investment and implementation complexity. On the other hand, fast-moving elements with shorter life cycles, such as agents and large language model (LLM) hosting, should be quick to implement and easy to change.

In this area, organizations can be flexible, first implementing the minimum necessary components for critical gen AI use cases, and then adding and removing components as needs evolve. For instance, a leading European bank implemented 14 key gen AI components across its enterprise architecture. This approach allowed the bank to implement 80 percent of its core gen AI use cases in just three months (Exhibit 1). By identifying the gen AI components with the largest potential impact early on, the bank focused its developer resources to produce gen AI features aligned with clear mid- to long-term goals. However, while a component-based approach to gen AI deployment is a crucial success factor for scaling gen AI, only 31 percent of gen AI high-performers and 11 percent of other companies have adopted this model.[3]

To succeed with a component-based gen AI development model, companies can create a task force to review, update, and evolve the road map. The task force also assigns execution plans, ensuring IT, data, AI, and business teams have appropriate responsibilities for specific rollouts. This requires clear communication between a variety of stakeholders, including AI engineers, software developers, data scientists, product managers, and enterprise architects, as well as regular reporting to business leads. Coordination is essential to ensure that component rollouts are systematized and aligned with organizational goals instead of presented in a series of disjointed pilots.

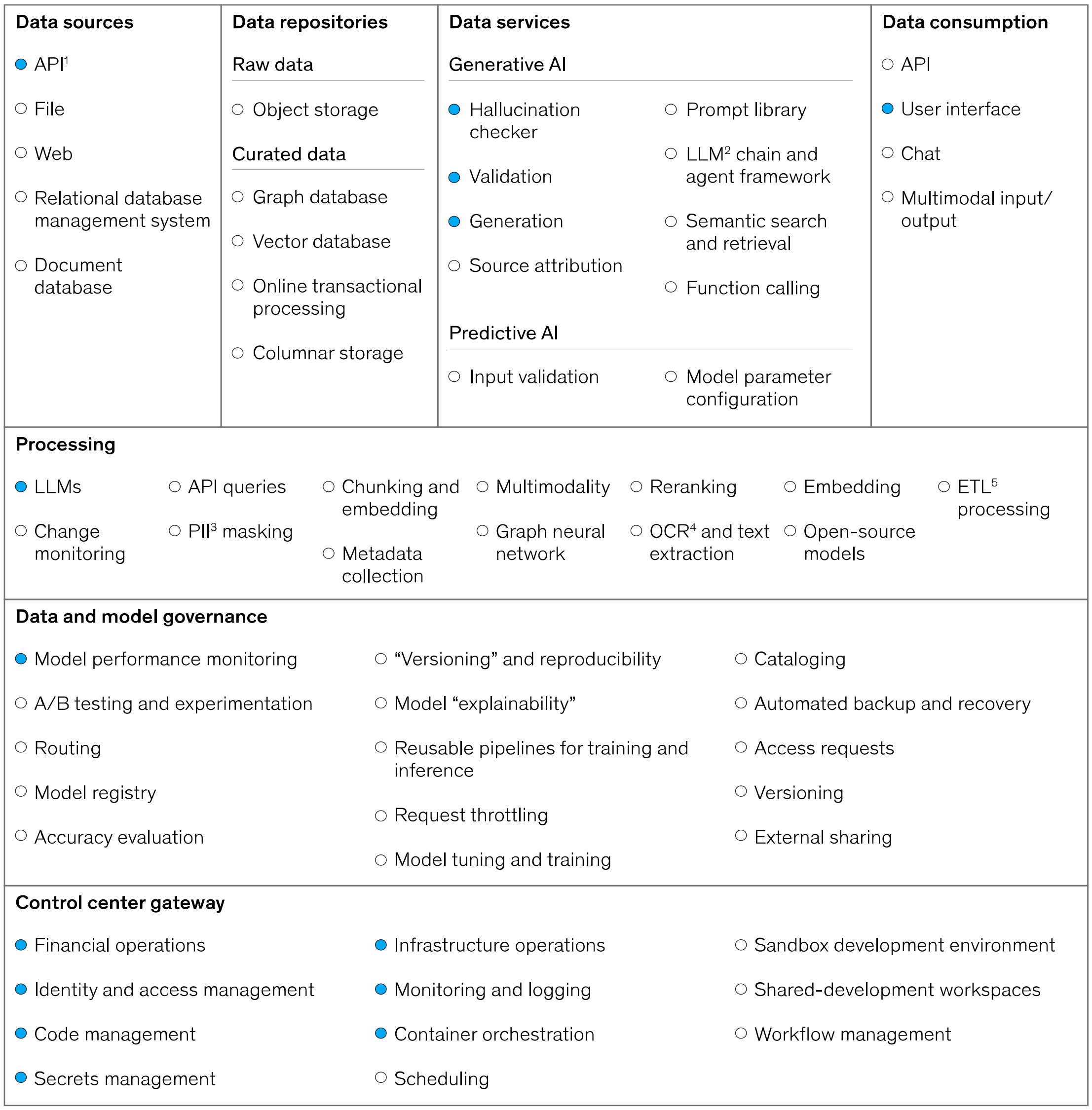
[3] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024.

Exhibit 1

Identifying core reusable components can allow an organization to roll out generative AI tools quickly.

Components for generative AI tools, illustrative

● Essential component



Data sources

- ● API[1]
- ○ File
- ○ Web
- ○ Relational database management system
- ○ Document database

Data repositories

Raw data

- ○ Object storage

Curated data

- ○ Graph database
- ○ Vector database
- ○ Online transactional processing
- ○ Columnar storage

Data services

Generative AI

- ● Hallucination checker
- ● Validation
- ● Generation
- ○ Source attribution
- ○ Prompt library
- ○ LLM[2] chain and agent framework
- ○ Semantic search and retrieval
- ○ Function calling

Predictive AI

- ○ Input validation
- ○ Model parameter configuration

Data consumption

- ○ API
- ● User interface
- ○ Chat
- ○ Multimodal input/output

Processing

- ● LLMs
- ○ Change monitoring
- ○ API queries
- ○ PII[3] masking
- ○ Chunking and embedding
- ○ Metadata collection
- ○ Multimodality
- ○ Graph neural network
- ○ Reranking
- ○ OCR[4] and text extraction
- ○ Embedding
- ○ Open-source models
- ○ ETL[5] processing

Data and model governance

- ● Model performance monitoring
- ○ A/B testing and experimentation
- ○ Routing
- ○ Model registry
- ○ Accuracy evaluation
- ○ "Versioning" and reproducibility
- ○ Model "explainability"
- ○ Reusable pipelines for training and inference
- ○ Request throttling
- ○ Model tuning and training
- ○ Cataloging
- ○ Automated backup and recovery
- ○ Access requests
- ○ Versioning
- ○ External sharing

Control center gateway

- ● Financial operations
- ● Identity and access management
- ● Code management
- ● Secrets management
- ● Infrastructure operations
- ● Monitoring and logging
- ● Container orchestration
- ○ Scheduling
- ○ Sandbox development environment
- ○ Shared-development workspaces
- ○ Workflow management

[1] Application programming interface. [2] Large language model. [3] Personally identifiable information. [4] Optical character recognition. [5] Extract, transform, load.

McKinsey & Company

Choose an extended or distinct gen AI team

When building a gen AI operating model, defining a core team is crucial. There are two main options: extend an existing data or IT team by equipping them with new gen AI skills or build a distinct and separate gen AI team. The latter can be accomplished by selecting people from an existing data or IT team or by hiring new talent. Each has its own advantages and constraints.

Making an existing data team responsible for gen AI may seem to be the easier option, though the pendulum could shift as gen AI matures. For instance, a leading logistics organization extended its IT organization, which included data teams, to launch several gen AI initiatives. The company wrapped gen AI into its data and analytics road map, encouraging existing teams to upskill in gen AI capabilities. While the company succeeded in deploying a gen AI pilot, it was limited in scope. And future rollouts were slower than expected because gen AI products were integrated into the company's overall technology platform, requiring time and resources to ensure compliance with existing systems.

Decoupling the gen AI team from the IT or data organization has different advantages. This approach allows an organization to build a new highly skilled gen AI team from scratch. With a solid foundation in data and AI architectures, the new team can quickly iterate on gen AI components outside of the larger IT function.

Several leading European banks have launched such gen AI task forces, with the idea they could potentially expand into full-fledged centers of excellence (CoE). In highly regulated industries such as healthcare and financial services, creating new, centralized gen AI teams also appears to be the best practice. Using this approach, these companies launched several gen AI projects within weeks instead of months.

Either model can be successful, but both have pitfalls companies should be careful to avoid.
If the gen AI team is decoupled from IT, its road map should still be aligned with the broader IT organization to avoid duplicating efforts or building disconnected gen AI components in multiple places. The capability map and ownership of each component should be clearly defined and shared across the organization. For example, the gen AI task force could oversee prompt engineering and guardrails, LLM operations and orchestration, and model improvement—but not data ingestion, management, and storage.

However, if the gen AI team expands as an offshoot of existing IT and data functions, the team will need to successfully manage two starkly different technology life cycles. Specific gen AI components, such as LLM hosting and model hubs, will need to be developed and put into production more rapidly than traditional IT and data components, such as hosting and containers.

Whether a company chooses an extended or distinct gen AI team, it is important for a central IT team to define a common underlying technology infrastructure that ties all gen AI tools together.

Avoiding this step could lead to compliance issues or technical debt—the extra work required to fix buggy products that were initially built for speed rather than quality.

Prioritize data management in strategic business domains

As every data leader knows, effective data management is a pivotal factor in implementing gen AI. Without a functional data organization, gen AI applications will not be able to retrieve and process the right information they need. Yet most enterprises report significant hurdles in data utilization, including issues with model reusability, accessibility, scalability, and quality. That is why a data management and governance strategy should be part of any operating model for gen AI. Governance includes managing document sourcing, preparation, curation, and tagging, as well as ensuring data quality and compliance, for both structured and unstructured data.

Managing vast amounts of unstructured data, which comprise more than 80 percent of companies' overall data, may seem like a daunting task.[4] Indeed, 60 percent of gen AI high performers and 80 percent of other companies struggle to define a comprehensive strategy for organizing their unstructured data.[5] To address this challenge, organizations can prioritize specific domains and subdomains of unstructured data based on business priorities. For example, one company may prioritize a data domain that groups all gen AI products under development into one business unit, whereas another may prioritize a domain that groups all data related to a specific function, such as finance or HR. The ideal domains and subdomains should be small enough to be actionable while being sufficiently large enough to provide a significant, measurable outcome.

Since handling unstructured data can be unfamiliar to many data teams, the process should be launched by experts in a centralized manner. These experts are typically data engineers trained to handle unstructured data, as well as natural-language-processing engineers, grouped into a CoE. They establish and implement processes for managing unstructured data so it is accessible to gen AI systems. They ensure policies in the company's gen AI operating model provide a view on when and where data is consumed. They also ensure consistent standards for data quality, risk management, and compliance.

However, once the CoE provides a deployment road map, domain experts with business oversight should take over the data management process. They are better equipped to extract knowledge from specific records in their field than data professionals alone. As business units begin to provide more higher-quality data for a wider variety of use cases, the centralized data teams tend to become overwhelmed by the demand and lack the expertise to check the quality, veracity, and tagging of domain-specific documents.

Plan for a decentralized approach to gen AI development

As domain teams become more adept at managing data, companies may choose to progressively increase these teams' ownership of gen AI development—moving from a centralized model to a federated one and finally to a decentralized one (Exhibit 2). Forward-thinking data executives may want to ensure their gen AI operating road maps include future scenarios of decentralized development. There are three approaches to consider.

[4] Tam Harbert, "Tapping the power of unstructured data," MIT Sloan School of Management, February 1, 2021.
[5] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024.

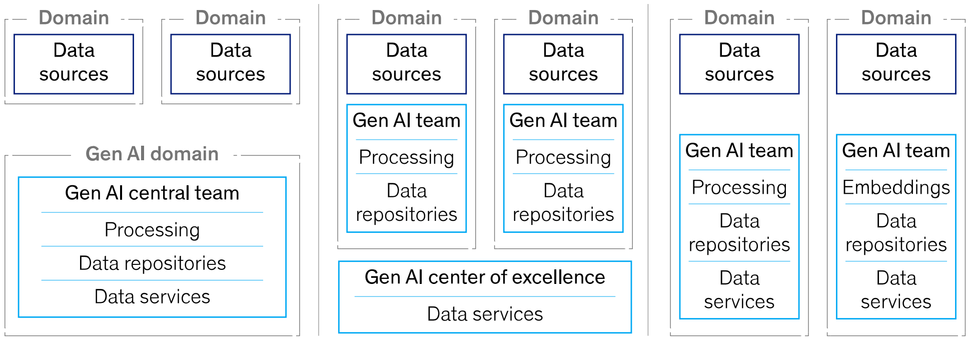
Centralized gen AI

Some companies choose to centralize gen AI into their own domains. This allows organizations to build capabilities quickly and control costs. A leading global telco used this model, making gen AI a node to its business units, operating under the leadership of a chief data and AI officer. The company was able to quickly set up a knowledgeable gen AI team by pulling existing employees into a central unit. This approach kept development costs low and reduced the risk of multiple teams creating similar projects.

Exhibit 2

Companies can start with centralized generative AI development and move to federated or decentralized approaches as deployments mature.

Archetypes of generative AI (gen AI) operating models, nonexhaustive



	Centralized	**Federated**	**Decentralized**
Prerequisites	• Agile way of working to integrate domain teams in squads, coordinated by a central team	• Sufficient domain maturity to drive processing and data repositories • Agile way of working to coordinate among central team and domains	• Very high domain maturity and sufficient scale to fully drive gen AI at domain level • Sufficient budget to build tech capabilities
Pros	• Faster skill and capability building thanks to concentration of talent • Higher control of costs	• Higher integration among center of excellence and domains, reducing friction and garnering support for use cases	• Higher buy-in from domain and specialized resources to drive relevant development • Ability to create agents directly for their needs
Cons	• Gen AI domain potentially siloed from key decision making because of distance from other domains	• Potential slowed execution in center of excellence, especially for cross-function use cases	• Potential lack of consistency, knowledge, and best practices • Higher cost and risk of duplicating efforts

McKinsey & Company

Federated gen AI
As companies build gen AI expertise, they often choose a federated model, in which business units are not only responsible for consuming data related to their domains but also take over data processing and repositories. This model allows domains to integrate gen AI more deeply into their daily workflows for stronger business outcomes.

A major North American investment bank chose the federated model to develop new gen AI use cases within a business unit. The gen AI use cases were so successful that the company later provided funding to scale similar gen AI tools across the organization. This lighthouse project model, in which an innovative project is developed within a business unit and then extended throughout the organization, can be a successful way to boost gen AI deployment without project duplication.

Decentralized gen AI
Some innovative organizations push decentralization even further, transferring all gen AI capabilities to domains. In this model, each domain creates its own gen AI team composed of business, data, and technical experts aligned on a common goal to develop relevant gen AI applications. A decentralized model of gen AI development allows domains to create gen AI agents tailored specifically to their needs, which they can then offer to other domains. For example, a marketing domain that creates agents for social media content creation and post management could then see those agents adopted by many other domains, such as business development, sales, and customer success. With this approach, it is important for a centralized IT team to retain visibility into the tools being developed to avoid blind spots and ensure no two teams develop similar gen AI tools.

Unify federated teams through common infrastructure

While business units know which everyday problems they need to solve with gen AI and are thus well placed to build specific use cases within their own domains, this decentralized development process should never compromise the company’s overall security or resiliency. Instead, companies should ensure IT teams build and manage an underlying common infrastructure on top of which all gen AI tools are developed and deployed. The IT team should also be responsible for building repeatable platforms that can be used by all the business units, such as a prompt library, a repository for Python code, standard agents, and systemized cloud storage. This type of centralized IT management empowers business units to create new gen AI tools, while ensuring all use cases they develop adhere to a highly secure and unified technology framework.

Emphasize risk and compliance governance

Gen AI comes with heightened risks, including potential hallucinations, misinformation, and data leaks. That is why every gen AI operating model should include explicit stipulations for risk and compliance governance. Companies can start by delineating the levels of risk they are willing to tolerate with gen AI and which areas of the business require more safeguards. This initial risk assessment evaluates the diverse ways a gen AI application could affect the company, customers, and partners. When this risk assessment is complete, companies can create a governance and monitoring plan, which should also define any new quantitative and qualitative tests that need to be conducted. By mitigating risks, companies can move forward with gen AI rollouts instead of taking a wait-and-see approach that could hamper competitiveness.

In practical terms, creating a gen AI risk plan involves a six-step process that data leaders must continually monitor and update as new potential risks arise and when new tools are deployed.

1. *Identify new risks:* Ask developers and technology users to identify potential AI-specific threats to add to the company's overall risk plan.
2. *Classify gen AI tools:* Encourage data teams to collaborate with the risk function to apply oversight to the most critical gen AI tools first.
3. *Deploy a tiered approach:* Adjust the depth and frequency of derisking methods for each gen AI tool based on continual risk assessments.
4. *Make risk tracking habitual:* Begin oversight at the development stage, continuing to measure risks throughout implementation and production.
5. *Equip risk teams for success:* Establish a CoE with developers and risk leaders to ensure the risk team keeps pace with evolving gen AI trends.
6. *Get everyone on board:* Ensure end users, developers, managers, and leaders all understand the company's policies for safe gen AI use.

By following the above guidelines, data leaders can establish a risk structure that balances oversight with the ability to support rapid decision making and agile gen AI deployments. A strong AI governance plan also helps companies keep pace with constantly shifting AI regulations. For example, the EU Artificial Intelligence Act emphasizes the need for transparency, requiring organizations to notify users about AI risks, ensure model output quality, and conduct regular compliance assessments. Legislation in many countries requires companies to meet privacy standards that can affect how gen AI tools are permitted to consume data. Any gen AI governance model must be flexible enough to take relevant regulations and updates into account.

As gen AI moves from experimentation to implementation, companies must create both operating and technical models to successfully guide deployments across their organizations. Data sits at the center of these models. Companies must organize all their data so gen AI applications can securely access it at scale. From an operational standpoint, data leaders should coordinate gen AI rollouts as genuine digital transformations, applying best practices to ensure clear governance, objectives and key results, and progress monitoring.

With such a coordinated plan, companies can quickly launch gen AI use cases from a centralized CoE. They can also prepare themselves for a longer-term evolution to a decentralized development approach supported by common technology infrastructure, which can power the agile gen AI deployments companies need to compete in today's fast-paced AI economy.

Alex Singla is a senior partner in McKinsey's Chicago office; **Asin Tavakoli** is a partner in the Düsseldorf office, where **Holger Harreis** is a senior partner; **Kayvaun Rowshankish** is a senior partner in the New York office; **Klemens Hjartar** is a senior partner in the Copenhagen office; and **Gaspard Fouilland** and **Olivier Fournier** are consultants in the Paris office.

The authors wish to thank Jean-Baptiste Dubois, Jon Boorstein, and Pedro J. Silva for their contributions to this article.