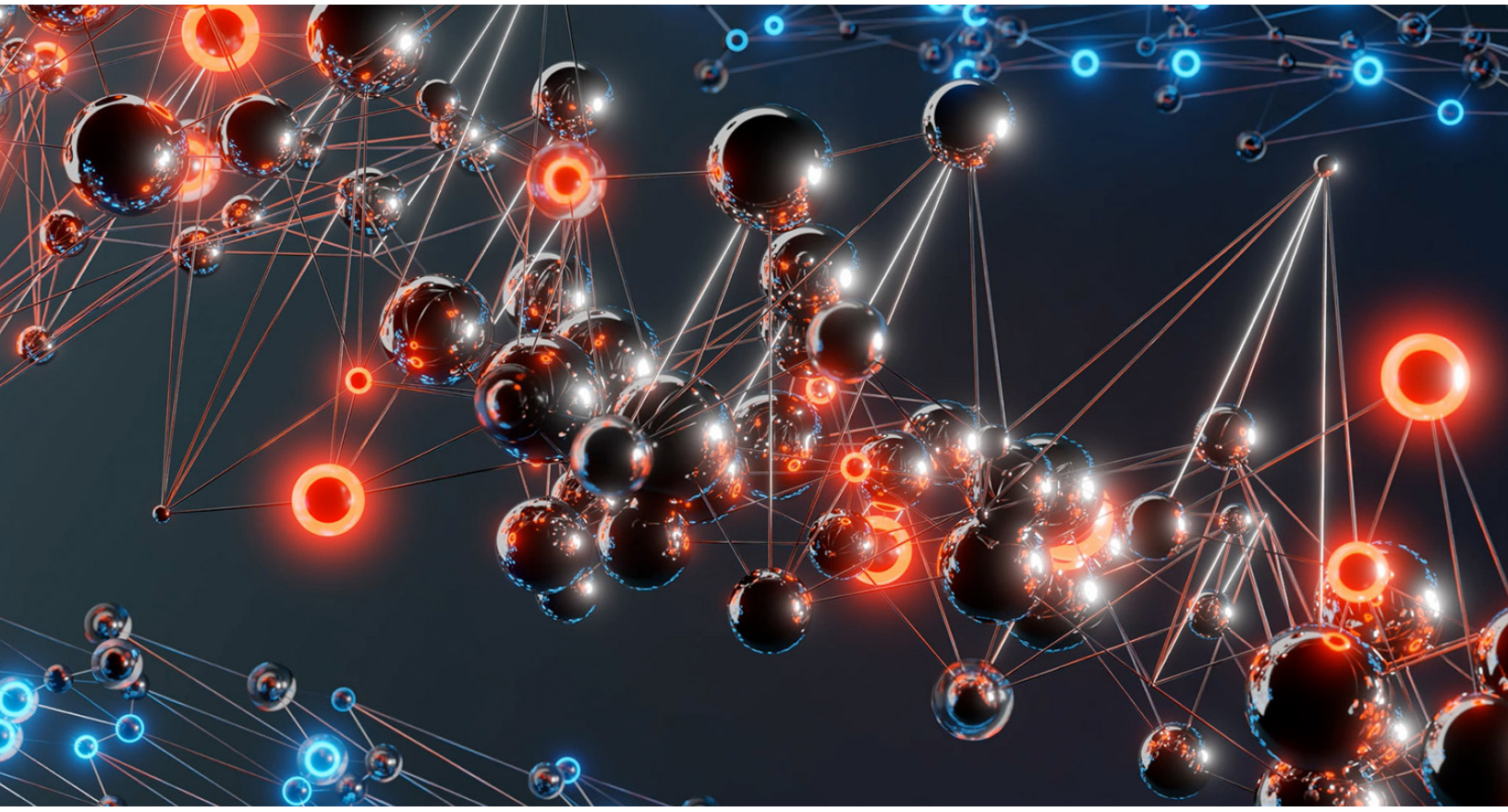
McKinsey & Company | QuantumBlack AI by McKinsey

Financial Services Practice

Scaling gen AI in banking: Choosing the best operating model

Generative AI is transforming financial services, offering opportunities for efficiency and innovation. As banks race to deploy gen AI, the right operating model can help unlock its potential.

This article is a collaborative effort by Kevin Buehler, Alison Corsi, Mina Jurisic, Larry Lerner, Andrea Siani, and Brian Weintraub, representing views from McKinsey's Banking Practice and Risk & Resilience Practice.



March 2024

Generative AI (gen AI) is revolutionizing the banking industry as financial institutions use the technology to supercharge customer-facing chatbots, prevent fraud, and speed up time-consuming tasks such as developing code, preparing drafts of pitch books, and summarizing regulatory reports.

The McKinsey Global Institute (MGI) estimates that across the global banking sector, gen AI could add between $200 billion and $340 billion in value annually, or 2.8 to 4.7 percent of total industry revenues, largely through increased productivity.[1] However, as banks and other financial institutions move to quickly implement the technology, challenges are emerging. Getting gen AI right can potentially unlock tremendous value; getting it wrong can lead to complications. Companies across industries face gen AI risks, including the generation of false or illogical information, intellectual property infringement, limited transparency in how the systems function, issues of bias and fairness, security concerns, and more.

In a previous article, we explored a series of strategies that banks could use to capture the full value of gen AI. Achieving sustained value, beyond initial proofs of concept, requires strong capabilities across seven dimensions:

— strategic road map

— talent

— operating model

— technology

— data

— risk and controls

— adoption and change management

These dimensions are interconnected and require alignment across the enterprise. A great operating model on its own, for instance, won't bring results without the right talent or data in place.

This article takes a closer look at one of these seven dimensions: the operating model, which is essentially a blueprint for how a business puts strategy into action. Subsequent articles will examine some of the other dimensions. In this article, we explain what an operating model is and why it is important, then delve into the operating-model archetypes that have emerged for gen AI in banking—including the one with the best record of success. Finally, we go over important decisions financial institutions need to make as they set up a gen AI operating model.

We have found that across industries, a high degree of centralization works best for gen AI operating models. Without central oversight, pilot use cases can get stuck in silos and scaling becomes much more difficult. Looking at the financial-services industry specifically, we have observed that financial institutions using a centrally led gen AI operating model are reaping the biggest rewards. As the technology matures, the pendulum will likely swing toward a more federated approach, but so far, centralization has brought the best results.

A centrally led gen AI operating model is beneficial for several reasons:

— Given the scarcity of top gen AI talent, centralization allows the enterprise to allocate talent in a way that is more likely to benefit the entire organization. A centrally led operating model can also help the organization build a world-class, cohesive gen AI team that fosters a sense of camaraderie, helping attract and retain talent.

— In a rapidly changing environment where new large language models and gen AI features are regularly being introduced, a central team can stay on top of the evolving gen AI landscape better than several teams dispersed across an organization.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.

— A centrally led operating model is useful early on in an enterprise's gen AI push, when it is necessary to make frequent and important decisions on matters such as funding, tech architecture, cloud providers, large language model providers, and partnerships.

— Risk management and keeping up with regulatory developments are easier with a centrally led approach.

Choosing an operating model isn't a simple binary approach, however. A financial institution can draw insights from the details explored in this article, decide how much to centralize the various components of its gen AI operating model, and tailor its approach to its own structure and culture. An organization, for instance, could use a centralized approach for risk, technology architecture, and partnership choices, while going with a more federated design for strategic decision making and execution.

The importance of the operating model

An operating model is a representation of how a company runs, including its structure (roles and responsibilities, governance, and decision making), processes (performance management, systems, and technology), and people (skills, culture, and informal networks).

Financial institutions that successfully use gen AI have made a concerted push to come up with a fitting, tailored operating model that accounts for the new technology's nuances and risks, rather than trying to incorporate gen AI into an existing operating model. We have observed that the majority of financial institutions making the most of gen AI are using a more centrally led operating model for the technology, even if other parts of the enterprise are more decentralized. This is likely to evolve as the technology matures.

The right operating model for a financial-services company's gen AI push should both enable scaling and align with the firm's organizational structure and culture; there is no one-size-fits-all answer. An effectively designed operating model, which can change as the institution matures, is a necessary foundation for scaling gen AI effectively.

In essence, a suitable operating model enables the financial institution to efficiently carry out three types of activities:

— *Strategic steering.* Identify clusters, or domains, of gen AI use cases that align with the enterprise's strategic objectives; sort them by priority into a road map that maximizes value while managing risk; and monitor value creation in order to ensure efficient resource allocation.

— *Standard setting.* Define common standards (such as those concerning technology architecture choices, data practices, and risk frameworks and controls) to increase efficiency and use insights learned from completed projects on new ones.

— *Execution.* Design and test use cases' technical solutions, put the use cases that meet the appropriate performance and safety criteria into production, and scale them if there is a business case for doing so, ensuring that their impact is tracked and delivered.

Operating-model archetypes for gen AI in banking

Banks and other financial institutions can take different approaches to how they set up their gen AI operating models, ranging from the highly centralized to the highly decentralized.

We recently conducted a review of gen AI use by 16 of the largest financial institutions across Europe and the United States, collectively representing nearly $26 trillion in assets. Our review showed that more than 50 percent of the businesses studied have adopted a more centrally led organization for gen AI, even in cases where their usual setup for data and analytics is relatively decentralized. This centralization is likely to be temporary, with the structure becoming more decentralized as use of

the new technology matures. Eventually, businesses might find it beneficial to let individual functions prioritize gen AI activities according to their needs.

Among the financial institutions we studied, four organizational archetypes have emerged, each with its own potential benefits and challenges (exhibit).

Highly centralized

Potential benefits. This structure—where a central team is in charge of gen AI solutions, from design to execution, with independence from the rest of the enterprise—can allow for the fastest skill and capability building for the gen AI team.

Potential challenges. The gen AI team can be siloed from the decision-making process. It can also be distant from the business units and other functions, creating a possible barrier to influencing decisions.

Centrally led, business unit executed

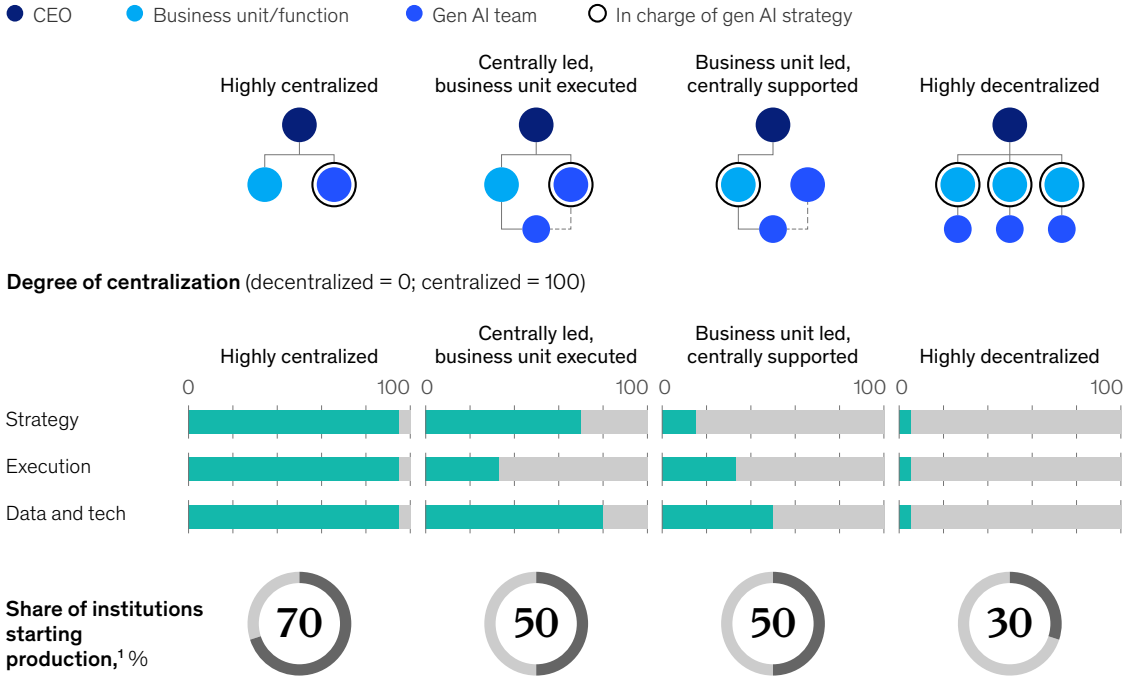
Potential benefits. This archetype has more integration between the business units and the gen AI team, reducing friction and easing support for enterprise-wide use of the technology.

Potential challenges. It can slow execution of the gen AI team's use of the technology because input

Exhibit

Four archetypes have emerged for using gen AI in financial services, and the highly centralized approach is showing the best results.

Organizational archetypes for generative AI operating models



[1]Share of financial institutions with this type of operating model that are putting gen AI use cases into production, moving beyond experimentation by having live use cases at the minimal-viable-product stage and beyond.

McKinsey & Company

and sign-off from the business units is required before going ahead.

Business unit led, centrally supported
Potential benefits. With this archetype, it is easy to get buy-in from the business units and functions, as gen AI strategies bubble from the bottom up.

Potential challenges. It can be difficult to implement uses of gen AI across various business units, and different units can have varying levels of functional development on gen AI.

Highly decentralized
Potential benefits. It is easy to get buy-in from the business units and functions, and specialized resources can produce relevant insights quickly, with better integration within the unit or function.

Potential challenges. Business units that do their own thing on gen AI run the risk of lacking the knowledge and best practices that can come from a more centralized approach. They can also have difficulty going deep enough on a single gen AI project to achieve a significant breakthrough.

The operating model with the best results

At this very early stage of the gen AI journey, financial institutions that have centralized their operating models appear to be ahead. About 70 percent of banks and other institutions with highly centralized gen AI operating models have progressed to putting gen AI use cases into production,[2] compared with only about 30 percent of those with a fully decentralized approach. Centralized steering allows enterprises to focus resources on a handful of use cases, rapidly moving through initial experimentation to tackle the harder challenges of putting use cases into production and scaling them. Financial institutions using more dispersed approaches, on the other hand, struggle to move use cases past the pilot stage.

The nascent nature of gen AI has led financial-services companies to rethink their operating models to address the technology's rapidly evolving capabilities, uncharted risks, and far-reaching organizational implications. More than 90 percent of the institutions represented at a recent McKinsey forum on gen AI in banking reported having set up a centralized gen AI function to some degree, in a bid to effectively allocate resources and manage operational risk.

Our surveys also show that about 20 percent of the financial institutions studied use the highly centralized operating-model archetype, centralizing gen AI strategic steering, standard setting, and execution. About 30 percent use the centrally led, business unit–executed approach, centralizing decision making but delegating execution. Roughly 30 percent use the business unit–led, centrally supported approach, centralizing only standard setting and allowing each unit to set and execute its strategic priorities. The remaining institutions, approximately 20 percent, fall under the highly decentralized archetype. These are mainly large institutions whose business units can muster sufficient resources for an autonomous gen AI approach.

Centralization isn't friction free. The main obstacles to implementing a centralized operating model have so far stemmed from disagreements over the strategic road map, funding mechanisms, and talent pooling as units fear losing out on crucial resources or having their operational priorities overlooked.

The financial-services companies that have best managed the transition to gen AI already had a high level of organizational agility, allowing them to quickly rework processes and flexibly pool resources, either by locating them in a central hub or by creating ad hoc, centrally coordinated, agile squads to execute use cases. Compared with a traditional AI squad, gen AI teams tend to feature more significant involvement from cloud engineers, business domain experts, and risk and compliance

[2] Live use cases at minimal-viable-product stage or beyond.

professionals from the beginning of a use case. This is because of two factors: the highly iterative nature of the gen AI development process and the need to consider, even in the early development stage, unforeseen or speculative implications of scaling the applications.

As gen AI technology and organizations' grasp of its implications mature, the operating model might swing toward a more federated design in both strategic decision making and execution, while standard setting is the likeliest candidate for continued centralization (for example, in risk management, tech architecture, and partnership choices).

A checklist of essential decisions to consider

Choosing and implementing a gen AI operating model requires leaders at financial institutions to make decisions in various areas, including both those directly implicated in the operating model and those that fall into other areas but affect how the model works. Here is a checklist executives can keep in mind as they come up with the best operating model for their organizations:

— *Strategy and vision.* First, the financial institution needs to decide which leaders will define its gen AI strategy and whether that will be done on an enterprise-wide or business unit level. This should include a vision for the potential value at stake and an assessment of which functions or processes are likely to be affected the most by gen AI.

— *Domains and use cases.* Next, the institution should ascertain who will determine the enterprise domains, or clusters, of gen AI use cases and the specific use cases within those domains.

— *Deployment model.* Regarding the implementation of the domains and use cases, the institution should decide whether it will be a "taker" (procuring targeted solutions from vendors), a "shaper" (integrating broader solutions from vendors), or a "maker" (developing in-house solutions that reshape the core business).

— *Funding.* The institution will need to set out how gen AI use cases will be funded, which will depend on how centralized or decentralized its gen AI approach is. Banks typically fund use cases through a combination of individual business units and a foundation-building central team dedicated to gen AI.

— *Talent.* The enterprise should define which skills will be needed for gen AI initiatives, then put in place the necessary talent through hiring, upskilling, strategic outsourcing, or a combination of all these strategies. Another step will be to determine the role of "translators" who understand both the business needs and technical requirements of implementing gen AI use cases and domains.

— *Risk.* The financial institution should determine who defines risk guardrails (such as those related to data privacy and intellectual property infringement) and mitigation strategies. It should also decide to what extent existing frameworks should be adjusted to account for risks specific to gen AI, including whether additional governance is required for particular use cases (such as customer-facing ones).

— *Change management.* A committee will need to lead the execution of a change management plan to ensure evolutions in mindsets and behaviors as required for the successful adoption of gen AI across the enterprise.

Without the right gen AI operating model in place, it is tough to incorporate enough structure and move quickly enough to generate enterprise-wide impact. To choose the operating model that works best, financial institutions need to address some important points, such as setting expectations for the gen AI team's role and embedding flexibility into the model so it can adapt over time. That flexibility pertains to not only high-level organizational aspects of the operating model but also specific components such as funding.

The dynamic landscape of gen AI in banking demands a strategic approach to operating models. Banks and other financial institutions should balance speed and innovation with risk, adapting their structures to harness the technology's full potential. As financial-services companies navigate this journey, the strategies outlined in this article can serve as a guide to aligning their gen AI initiatives with strategic goals for maximum impact. Scaling isn't easy, and institutions should make a push to bring gen AI solutions to market with the appropriate operating model before they can reap the nascent technology's full benefits.

Kevin Buehler is a senior partner in McKinsey's New York office, where **Alison Corsi** is a consultant, and **Brian Weintraub** is a partner; **Mina Jurisic** is a partner in the Paris office, where **Andrea Siani** is a consultant; and **Larry Lerner** is a partner in the Washington, DC, office.

The authors wish to thank Antonio Castro for his contributions to this article.

Designed by McKinsey Global Publishing