McKinsey Digital

Enterprise technology's next chapter: Four gen AI shifts that will reshape business technology

How tech leaders organize and manage teams, talent, IT architecture, and costs may dramatically shift in the next decade as a result of generative AI. Here's what to consider.

by James Kaplan, Mark Gu, and Megha Sinha

December 2024

Companies often overestimate the impact of short-term changes in technology and underestimate the effect of long-term changes. This well-known dynamic is particularly relevant for generative AI (gen AI) in enterprise technology. Today's many bold predictions about its impact on enterprise technology often focus on shorter-term horizons (with immediate focus on efficiency and productivity in two to three use cases) rather than on more forward-looking shifts and implications.

Our recent discussions with tech leaders across industries suggest that four emerging shifts are on the horizon as a result of gen AI, each with implications for how tech leaders will run their organizations. These include new patterns of work, architectural foundations, and organizational and cost structures that change both how teams interact with AI and the role gen AI agents play.[1]

A lot of work is still needed to enable this ambition. Only 30 percent of organizations surveyed earlier this year said they use gen AI in IT and software engineering and have seen significant quantifiable impact.[2] Moreover, organizations will need to understand and address the many risks of gen AI—including security, privacy, and explainability[3]—in order to take advantage of the opportunities.[4] But tech leaders we spoke with indicated that their organizations are already laying the groundwork.

From tools that support teams to AI 'artisan' and 'factory' teams

Conversations with tech leaders suggest a fundamental evolution is under way in how tech teams work, extending far beyond today's use of gen AI tools to increase individual productivity. Instead, leaders will restructure entire processes and workflows within every enterprise technology domain to integrate human and AI teams and enable peak human productivity and creativity. Teams may evolve with two new human–AI patterns of interaction: the "factory" and the "artisan" (Exhibit 1).

- **Factory pattern.** In this model, leaders deploy autonomous gen AI–enabled agents that can collaborate and navigate the work end to end. This approach is best for predictable, routine processes within enterprise technology, such as log monitoring, regulatory technology updates, or legacy code migration.[5] We've seen some promising early results from such an approach, with organizations modernizing code in nearly half the time by orchestrating gen AI agents to handle rote coding activities.[6]

- **Artisan pattern:** In this model, gen AI tools are implemented at scale to serve as assistants, aiding and enhancing the work of experienced software engineers and enterprise technology strategists and executives. This approach is most appropriate for processes that require human judgment and ingenuity, such as enterprise technology cost management and optimization and vendor sourcing and evaluation. These are non-deterministic activities, meaning there may be many potential ways to solve the problem, which makes them generally poor candidates for an autonomous AI agent to handle because they require higher-level thinking and often lack the large volumes of training data necessary for large language models (LLMs) to arrive at a well-determined answer.

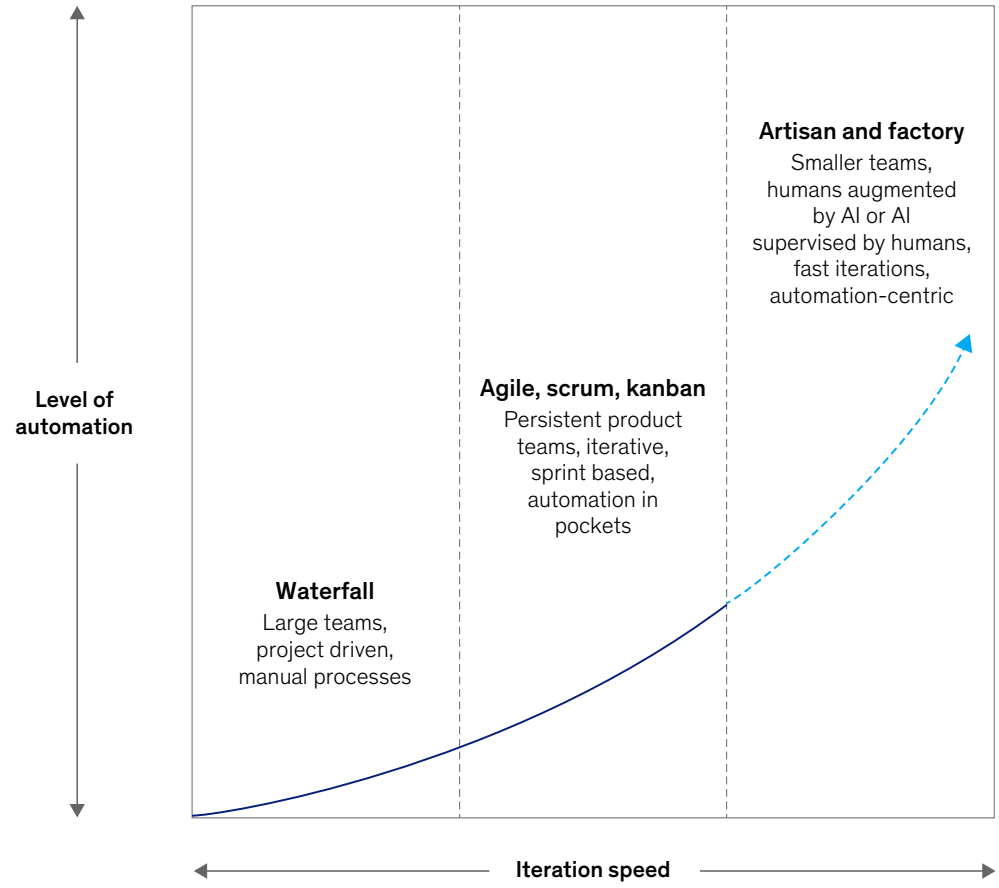
Leaders we spoke with agree that the right human–AI team model (artisan versus factory) will vary by enterprise technology domain and use case. For instance, many suggested that user acceptance testing will likely continue to be primarily human-led, with some automation, while the reverse would

[1] "The promise and the reality of gen AI agents in the enterprise," McKinsey, May 17, 2024.
[2] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024.
[3] Carlo Giovine and Roger Roberts, with Mara Pometti and Medha Bankhwal, "Building AI trust: The key role of explainability," McKinsey, November 26, 2024.
[4] "Implementing generative AI with speed and safety," *McKinsey Quarterly,* March 13, 2024.
[5] Aamer Baig, Sven Blumberg, Arun Gundurao, and Basel Kayyali, "Breaking technical debt's vicious cycle to modernize your business," McKinsey, April 25, 2023.
[6] "AI for IT modernization: Faster, cheaper, better," McKinsey, December 2, 2024.

Exhibit 1

Teams may evolve with two new human–AI interaction patterns.

Iteration speed vs level of automation (illustrative)



McKinsey & Company

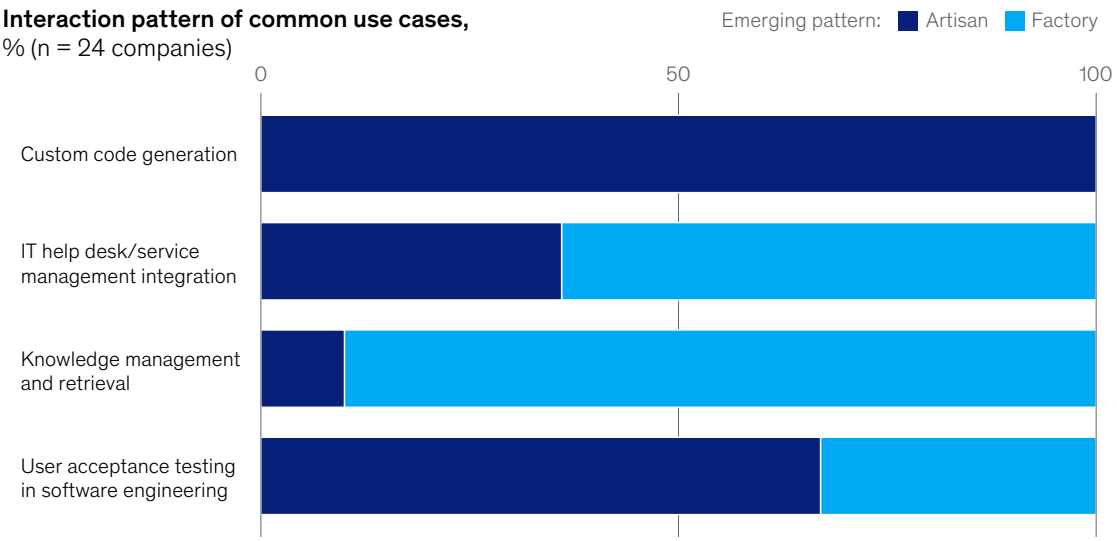
be true for IT service management (Exhibit 2). One of the challenges leaders may face is effectively blending these approaches to create a fluid, synchronized workflow as tasks move from human to AI and back to human. To support this work, tech leaders should consider developing a framework and governance and risk management strategy that guides their efforts.[7] What work should be AI-led, and what should be human-led? How do you effectively combine factory and artisan approaches to optimize enterprise technology delivery? What work remains unchanged because AI provides little value beyond a rudimentary productivity boost? How do you manage the handoffs from human to AI (and vice versa) within each process? Each workflow will need to be evaluated, deconstructed, and reconstructed with clearly defined rules of conduct.

New roles and skills will also be necessary. For AI-led tasks, "factory" supervisors will be needed who can oversee and implement audit mechanisms, validate AI output samples, identify and correct any AI agent deviations from the intended behavior,

[7] "Implementing generative AI with speed and safety," March 13, 2024.

Exhibit 2

How enterprise technology teams interact with gen AI will vary based on use case.



Source: McKinsey interviews on gen AI for enterprise technology

McKinsey & Company

and ensure the necessary governance and explainability are in place to maintain a high level of accuracy and trust. Where processes are human-led, enterprise technology experts will need to broaden their strategy, judgment, and execution skills—for example, by rapidly iterating on solutions to a problem rather than handing work off to junior staff, or by leveraging ideas provided by gen AI to activate brainstorming and inspire new thinking.

As the adoption of factory and artisan patterns scales from a few enterprise technology domains to many, the amount of technical debt (which accounts for about 40 percent of IT balance sheets), along with the time and cost of managing the environment, is expected to decrease significantly.[8] As a result, staff focused on daily maintenance can be reallocated to innovation.

This can trigger two important changes. First, the scale of innovation is expected to increase as IT directs more resources to strategic problem solving, closing out the backlog of business requests, and taking on new requests from the business. Second, as staff productivity increases, the speed at which IT can conceive, build, and launch capabilities is also expected to increase,[9] while the cost of that work will likely decrease. Gen AI may not reduce enterprise technology budgets outright. Instead, it will drive a strategic reallocation within the enterprise technology portfolio, with tech leaders increasingly focusing on growth-oriented projects rather than routine maintenance.

Tech leaders will likely need to strengthen planning and risk management efforts to sustain the rapid pace. It may involve, for example, aligning with

[8] "Breaking technical debt's vicious cycle to modernize your business," April 25, 2023.
[9] "Gen AI and beyond: Where else to focus now," McKinsey, July 12, 2024.

the C-suite much more frequently on innovation road maps to focus on value over volume amid a plethora of new business requests.[10] Rigorous guardrails for explainability, security, privacy, and other AI risks should be designed to enable both speed and scale in a safe and responsible way.[11] Tech leaders should also consider regularly reviewing how their organizations apply both artisan and factory patterns and fine-tune their approach to effectively balance cost-efficiency and innovation as business priorities evolve. This can include exploring additional opportunities to apply autonomous agents within workflows as well as ways to increase the productivity of artisan teams.

From application architectures dominating the landscape to predominantly AI agent and data architectures

IT architectures are also expected to be significantly different, evolving from a traditional application-focused approach to new multiagent architectures[12] where tech leaders oversee hundreds or thousands of distinct gen AI agents that can communicate with one another and the outside world to achieve a common goal.[13] For example, a fleet of gen AI agents might interact with existing inventory, supply chain, and analytics systems to automatically monitor stock levels, identify low inventory, and generate and send purchase orders to suppliers without complex integrations.

Tech leaders are expected to deploy these agents within their environments in three primary ways (Exhibit 3):

— **Super platforms.** Super platforms represent the next generation of third-party business applications—such as collaboration tools, customer relationship management (CRM), or enterprise resource planning (ERP) solutions—with built-in gen AI agents. These agents are essentially commodities that can be rapidly put into service. A CRM super platform, for instance, could enable a user to not only run a sales report but automatically communicate with the company's analytics tools without doing any programming.

— **AI wrappers.** These are essentially intermediary platforms that enable enterprise services to communicate and collaborate with third-party services via APIs without exposing their proprietary data. A bank, for example, may build a gen AI–enabled wrapper around an internally developed AI-driven credit risk model. The AI wrapper can then initiate any queries—for example, leveraging a vendor's LLM to generate risk factors based on the bank's customer data and credit scores. Once built, AI wrappers can easily interact with any third-party service, enabling IT to easily switch vendors when needed.

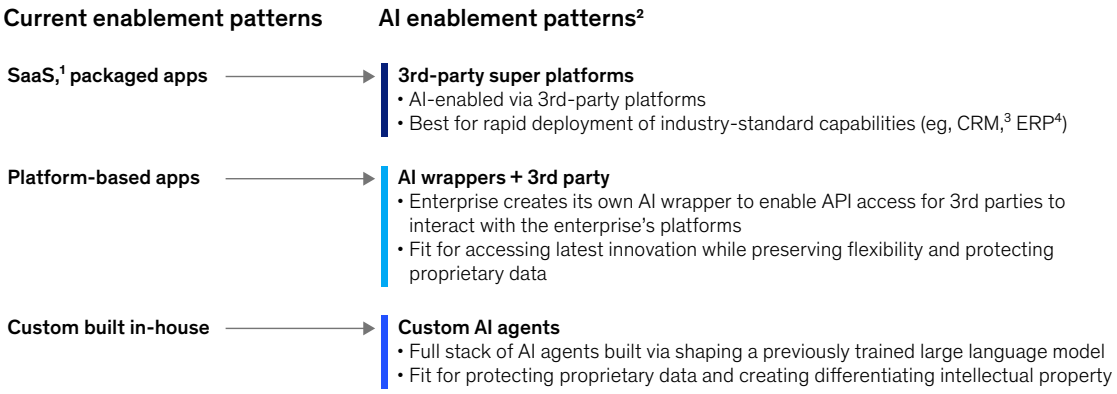
— **Custom AI agents.** Custom gen AI–enabled agents are internally developed by fine-tuning a pretrained LLM or using retrieval-augmented generation (RAG) with a company's proprietary data.[14] For example, enterprise technology may feed an existing model with customer data, call-center transcripts, company policies, and other internal information to create a gen AI agent that can assist call-center staff in responding to customer questions.

Which platform strategy an enterprise technology organization chooses may depend on a number of factors, including the potential of proprietary data to competitively differentiate the business. With super platforms, organizations give vendors access to their proprietary data, which may be anonymized but will likely be used to train

[10] Eric Lamarre, Kate Smaje, and Rodney Zemmel, "Rewired to outcompete," *McKinsey Quarterly,* June 20, 2023.
[11] "Implementing generative AI with speed and safety," March 13, 2024.
[12] "AI for IT modernization," December 2, 2024.
[13] Lareina Yee, Michael Chui, and Roger Roberts, with Stephen Xu, "Why agents are the next frontier of generative AI," *McKinsey Quarterly,* July 24, 2024.
[14] "Technology's generational moment with generative AI: A CIO and CTO guide," McKinsey, July 11, 2023.

Exhibit 3

Architectures may evolve with three AI enablement patterns.



[1]Software as a service.
[2]For more on strategic gen AI considerations, see "Technology's generational moment with generative AI: A CIO and CTO guide," McKinsey, July 11, 2023.
[3]Customer relationship management.
[4]Enterprise resource planning.

McKinsey & Company

and further improve the model to benefit the vendor's customers, including competitors. As a result, domains with sensitive or proprietary data may best be served by securing the data inside an internal platform using an AI wrapper. Understanding when to safeguard proprietary data to maintain competitive advantage will be an ongoing concern for tech leaders, because the answer is not always obvious.

Moreover, designing and managing a multiagent architecture that effectively leverages different agentic systems requires fundamentally different considerations than managing application-centric architectures. Modular frameworks that guide the development of common reusable gen AI agents and tools are important to ensure agents can be easily modified and assembled, like LEGOs, for use in different agentic workflows (Exhibit 4). Architectures should include agents designed to manage incoming requests and facilitate work across all steps of a process, such as orchestration agents for coordinating tasks across multiple agents, or communicator agents that can share messages and updates with agents throughout a workflow. They should also include task-oriented agents, such as a planner agent that can determine what steps are needed to complete a given request or a research agent that gathers and synthesizes information from diverse sources to support decision making.

A central enterprise technology team may provide the overall platform, including a data layer for consistent high-quality training data for all use cases; tools and apps, such as an image generator or API, that enable agents to interact with different third-party and enterprise systems; and risk controls. Data science and business teams could provide agents to address specific, business-critical problems. This approach allows business units to move quickly by directly building and experimenting with agents.

As a result, the primary priority of enterprise technology shifts from traditional release cycles and updates to continuous improvement of gen AI agents and the underlying data sets. Robust feedback mechanisms are an important part of this process because they enable enterprise technology organizations to improve agent performance as well as ensure efficient collaboration and orchestration of tasks across all agents in their ecosystem.

Exhibit 4

A new functional architecture is required for agentic workflows.

Illustrative consumption layer (business use cases)

1 Agentic platform: Agents, agent templates, tools for each agent, system prompts, and routing capabilities to help decide which agent to use are configured using the platform for a particular workflow

Agentic platform

2 Agents: Execute the step with the preconfigured skill and send the response back to orchestrator agent; these agents can be either generative or deterministic

Planner				
Visualizer	Researcher	Critic	Recommender	Sales expert
Knowledge searcher	Chat agent	Email author	Coder	Scheduler

3 Agent template: Entry point for all questions; incoming execution is coordinated by router agent and then passed to planner agent to coordinate steps

Orchestrator	Router	Memory	Feedback	Decorator
Prompt library	Communicator	Caching	Dashboard	…

4 Tools and apps: All custom-specific tools, such as API integrations, are consumed by the agents

Email hub	HTML writer	Calendar API	PPT[1] writer
Image generator	Metrics	Chart generator	

5 Data layer: Enterprise-level data services designed as systems of record for all use cases

Data enrichment services	Data generation services	Data preprocessing services
Fine-tuning services	RAG[2] services	Topic and attribute extraction services

[1]PowerPoint presentation.
[2]Retrieval-augmented generation.

McKinsey & Company

From a 'pyramid' or 'diamond' organizational structure to a flatter one, with new workforce development considerations

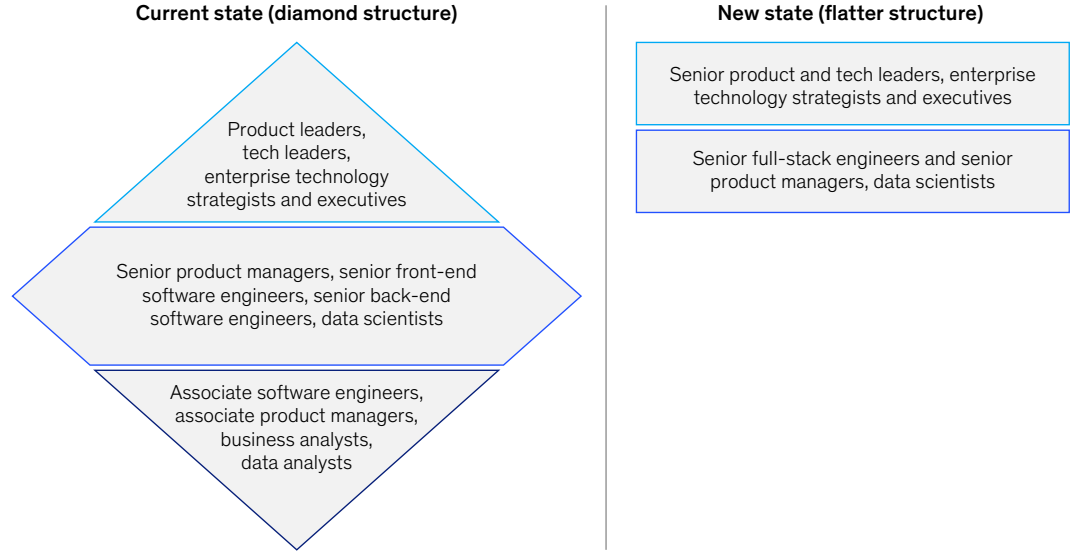
As automation and AI–human collaboration scales within IT, tech leaders will likely begin reshaping their organizational structure to capture the full productivity benefit (Exhibit 5).[15] Midlevel employees in domains most relevant for the artisan team model—such as software development, user experience design, and IT finance—will increasingly assume more integrated roles spanning the full scope of strategy and execution, enabling them to test and fine-tune ideas with business experts in real time. For domains most relevant for the factory team model—such as IT operations and help desk—we will see a significant flattening of the organization and thinning of junior roles, as well as a need for a few supervisors who can oversee and implement

[15] "Generative AI and the future of work in America," McKinsey Global Institute, July 26, 2023.

Exhibit 5

Flatter organizational structures will emerge.

Illustrative software development example



McKinsey & Company

audit mechanisms, validate AI output samples, and identify and correct any AI agent deviations from the intended behavior. In software development,[16] for example, full-stack engineers, proficient in business strategy and AI-enabled development, will drive upstream tasks such as understanding end-user intent and business outcomes, conceptualizing and developing a prototype of new product functionality with product managers and business leaders, and developing code frameworks (for instance, code libraries and support programs). For some projects, such as the development of an internal AI-powered analytics dashboard for a company's sales or marketing team, full-stack engineers empowered by gen AI may take on combined roles (their own and the product manager's role), leading both business ideation and product development.

Communication is a critical emerging skill, and it's needed to ensure that engineers can more effectively engage with leaders, peers, and customers. With AI automating many basic tasks and elevating staff productivity and capabilities, the roles of senior and junior experts may dramatically change. In software development, for example, tech leads and product managers may shift to new roles that draw on their experience to solve highly complex business problems or provide expertise in supervising gen AI agents. Junior developers may be asked to oversee distinct agents rather than writing code. A flatter and potentially leaner enterprise technology organizational model may emerge as a result.

This will undoubtedly result in a skills revolution with new workforce and career development considerations. How do you tailor upskilling and learning journeys to prepare staff for these changes?[17] How do you train senior experts for new roles that require them to break down complex problems, challenge assumptions (and gen AI agent decisions), and identify creative solutions with little data at hand? How do you create a

[16] "The gen AI skills revolution: Rethinking your talent strategy," *McKinsey Quarterly,* August 29, 2024.
[17] "Technology's generational moment with generative AI," July 11, 2023.

strong pipeline of enterprise technology cost-management experts, operations specialists, and full-stack engineers when much of the work that junior staff have traditionally cut their teeth on is now automated? How do you train staff to fully leverage gen AI in these new ways, but not use it as a crutch in place of critical and creative thinking? Beyond training, new recruitment and partnerships with external vendors will likely be needed to acquire specialized AI expertise where needed, including prompt engineering, which will be crucial in optimizing compute costs.

Additionally, before organizations make this shift, developing a clear baseline and process to measure impact from gen AI can help guide workforce planning and upskilling decisions. Ideally, metrics should track individual productivity (based on where and how time is spent), development speed (such as the cycle time from software product ideation to launch), and quality.

From application- to infrastructure-based cost structures, with increased focus on compute spend

As staff productivity rises, enterprise technology labor costs will decrease. At the same time, tech leaders may spend more on infrastructure, specifically compute, to support gen AI agents—a shift that will demand close attention to reduce the risk of compute costs outpacing the productivity gains generated from gen AI.

In optimizing compute spend, leaders should carefully examine compute and storage expenses across the entire life cycle, from initial planning to ongoing management of new gen AI agents.

As leaders evaluate which activities should be AI-led, measuring the cost of workflows with and without gen AI agents can provide valuable insight. Some processes may be performed less expensively over the long term with a mix of onshore and offshore talent after considering the computational and operational expenses (including upskilling and governance costs) necessary to produce accurate outputs from agents.

During development of gen AI agents, conducting cost-optimization reviews can surface the most cost-efficient options. For example, smaller, purpose-built LLMs may require less compute than large LLMs to achieve the same level of predictive and output accuracy necessary for a given process. More-precise prompts or dynamic resource allocation may also help trim some costs from the bottom line. Early learnings from our work suggest that costs can add up quickly when using AI models and approaches that haven't been optimized compared with those that have.

Finally, once gen AI agents are in operation, leaders should consider monitoring compute spend on an ongoing basis to avoid runaway costs. Many issues—including higher-than-expected use of agents, more complex queries than tested, and unresolved task dependencies that cause agents to repeat steps—may result in significantly higher compute costs than intended. As with cloud, using a FinOps as code (FaC) approach to integrate financial management into factory operations can provide the real-time insight leaders need to proactively identify issues and take action to manage operational costs effectively.

Full adoption of this new enterprise technology operating model is likely a decade away, and success will require more than tooling. It hinges on understanding where to implement factory and artisan patterns, designing an effective agent architecture, and preparing for the many implications on talent, cost, operations, and risk. Starting with a few enterprise technology domains can help leaders build their organizational muscle for operating in these new ways and extrapolate learnings to gain efficiencies as they scale. Given the scale of change, the journey will be challenging, yet the long-term impact will likely be greater than currently perceived.

James Kaplan is a partner in McKinsey's New York office, **Mark Gu** is a partner in the Boston office, and **Megha Sinha** is a partner in the New Jersey office.

The authors wish to thank Dhruv Pillai, Eva Li, Rob Kruszeski, and Stephen Robinson for their contributions to this article.