Seizing the agentic AI advantage

A CEO playbook to solve the gen AI paradox and unlock scalable impact with AI agents

Alexander Sukharevsky
Dave Kerr
Klemens Hjartar
Lari Hämäläinen
Stéphane Bout
Vito Di Leo
Guillaume Dagorret

June 2025

Contents

Foreword

by Arthur Mensch, CEO of Mistral AI

We're at a moment when gen AI has entered every boardroom, but for many enterprises, it still lingers at the edges of actual impact. Many CEOs have greenlit experiments, spun up copilots, and created promising prototypes, but only a handful have seen the needle move on revenue or impact. This report gets to the heart of that paradox: broad adoption with limited return.

The current diagnosis is this: Today, AI is bolted on. But to deliver real impact, it must be integrated into core processes, becoming a catalyst for business transformation rather than a sidecar tool. Most deployments today use AI in a shallow way—as an assistant that sits alongside existing workflows and processes—rather than as a deeply integrated, engaged, and powerful agent of transformation.

Agentic AI is the catalyst that can make this transition possible, but doing so requires a strategy and a plan to successfully power that transformation. Agents are not simply magical plug-n-play pieces. They must work across systems, reason through ambiguity, and interact with people—not just as tools, but as collaborators. That means CEOs must ask different questions: not "How do we add AI?" but "How do we want decisions to be made, work to flow, and humans to engage in an environment where software can act?"

Redefining how decisions are made, how work is done, and how humans engage with technology requires alignment across goals, tools, and people. That alignment can only happen when openness, transparency, and control are central to your technology and implementation—when builders have an open, extensible, and observable infrastructure and users can easily craft and use agents with the confidence that the work of agents is safe, reliable, and under their control. That alignment creates the trust and effectiveness that is the currency of scalable transformation that delivers results rather than regrets.

The technology to build powerful agents is already here. The opportunity now is to deploy agents in ways that are deeply tied to how value is created and how people work. That requires an architecture that is modular and resilient and, more importantly, an operating model that centers on humans—not just as users but as co-architects of the systems they will be living and working with.

This report lays out the playbook not for tinkering but for reinvention. ROI comes from strong intent: define the outcomes, embed agents deep in core workflows, and redesign operating models around them. Organizations that win will pair a clear strategy with tight feedback loops and disciplined governance, using agents to rethink how decisions are made and how work gets done—and turning novelty into measurable value.

At a glance

- Nearly eight in ten companies report using gen AI—yet just as many report no significant bottom-line impact.[1] Think of it as the "gen AI paradox."

- At the heart of this paradox is an imbalance between "horizontal" (enterprise-wide) copilots and chatbots—which have scaled quickly but deliver diffuse, hard-to-measure gains—and more transformative "vertical" (function-specific) use cases—about 90 percent of which remain stuck in pilot mode.

- AI agents offer a way to break out of the gen AI paradox. That's because agents have the potential to automate complex business processes—combining autonomy, planning, memory, and integration—to shift gen AI from a reactive tool to a proactive, goal-driven virtual collaborator.

- This shift enables far more than efficiency. Agents supercharge operational agility and create new revenue opportunities.

- But unlocking the full potential of agentic AI requires more than plugging agents into existing workflows. It calls for reimagining those workflows from the ground up—with agents at the core.

- A new AI architecture paradigm—the agentic AI mesh—is needed to govern the rapidly evolving organizational AI landscape and enable teams to blend custom-built and off-the-shelf agents while managing mounting technical debt and new classes of risk. But the bigger challenge won't be technical. It will be human: earning trust, driving adoption, and establishing the right governance to manage agent autonomy and prevent uncontrolled sprawl.

- To scale impact in the agentic era, organizations must reset their AI transformation approaches from scattered initiatives to strategic programs; from use cases to business processes; from siloed AI teams to cross-functional transformation squads; and from experimentation to industrialized, scalable delivery.

- Organizations will also need to set up the foundation to effectively operate in the agentic era. They will need to upskill the workforce, adapt the technology infrastructure, accelerate data productization, and deploy agent-specific governance mechanisms. The moment has come to bring the gen AI experimentation chapter to a close—a pivot only the CEO can make.

[1] "The state of AI: How organizations are rewiring to capture value," McKinsey, March 12, 2025.

About QuantumBlack, AI by McKinsey

QuantumBlack, McKinsey's AI arm, has been helping businesses create value from AI since 2009, expanding on McKinsey's technology work over the past 30 years. QuantumBlack combines an industry-leading tech stack with the strength of McKinsey's 7,000 technologists, designers, and product managers serving clients in more than 50 countries. With innovations fueled by QuantumBlack Labs—its center for R&D and software development—QuantumBlack delivers the organizational rewiring that businesses need to build, adopt, and scale AI capabilities.

1

The gen AI paradox: Widespread deployment, minimal impact

Key points:

- Nearly eight in ten companies have deployed gen AI in some form, but roughly the same percentage report no material impact on earnings.[2] We call this the "gen AI paradox."

- The main issue is an imbalance between "horizontal" and "vertical" use cases. The former, such as employee copilots and chatbots, have been widely deployed but deliver diffuse benefits, while higher-impact vertical, or function-specific, use cases seldom make it out of the pilot phase because of technical, organizational, data, and cultural barriers.

- Unless companies address these barriers, the transformational promise of gen AI will remain largely untapped.

Gen AI is everywhere—except in company P&L

Even before the advent of gen AI, artificial intelligence had already carved out a key place in the enterprise, powering advanced prediction, classification, and optimization capabilities. And the technology's estimated value potential was already immense—between $11 trillion and $18 trillion globally[3]—mainly in the fields of marketing (powering capabilities such as personalized email targeting and customer segmentation), sales (lead scoring), and supply chain (inventory optimization and demand forecasting). Yet AI was largely the domain of experts. As a result, adoption across the rank and file tended to be slow. From 2018 to 2022, for example, AI adoption remained relatively stagnant, with about 50 percent of companies deploying the technology in just one business function, according to McKinsey research (Exhibit 1).

Gen AI has extended the reach of traditional AI in three breakthrough areas: information synthesis, content generation, and communication in human language. McKinsey estimates that the technology has the potential to unlock $2.6 trillion to $4.4 trillion in additional value on top of the value potential of traditional analytical AI.[4]

Two and a half years after the launch of ChatGPT, gen AI has reshaped how enterprises engage with AI. Its potentially transformative power lies not only in the new capabilities gen AI introduces but also in its ability to democratize access to advanced AI technologies across organizations. This democratization has led to widespread growth in awareness of, and experimentation with, AI: According to McKinsey's most recent Global Survey on AI,[5] more than 78 percent of companies are now using gen AI in at least one business function (up from 55 percent a year earlier).

However, this enthusiasm has yet to translate into tangible economic results. More than 80 percent of companies still report no material contribution to earnings from their gen AI initiatives.[6] What's more, only

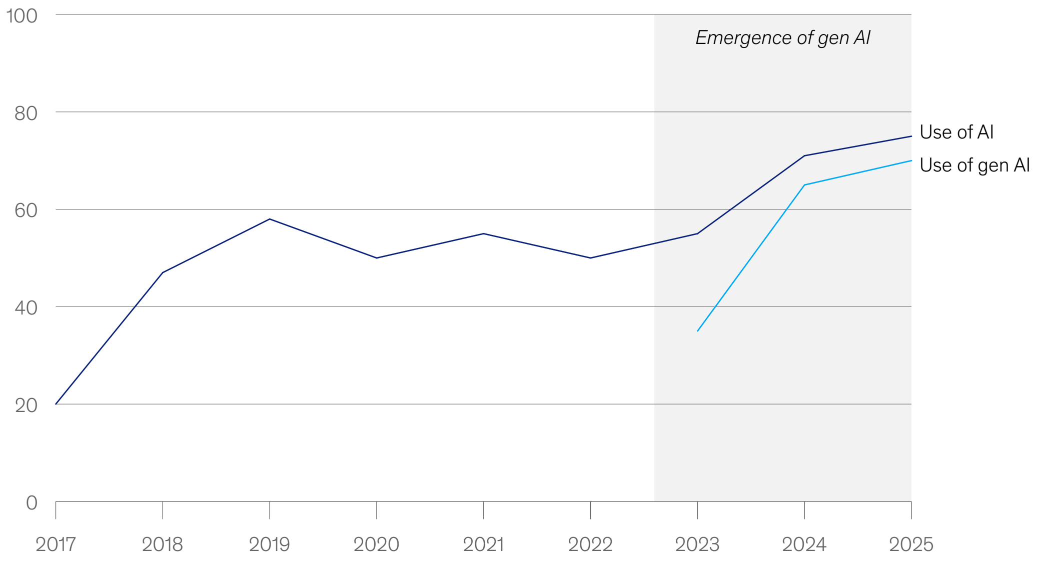
[2] "The state of AI: How organizations are rewiring to capture value," McKinsey, March 12, 2025.
[3] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[4] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[5] "The state of AI: How organizations are rewiring to capture value," McKinsey, March 12, 2025.
[6] "The state of AI: How organizations are rewiring to capture value," McKinsey, March 12, 2025.

1 percent of enterprises we surveyed view their gen AI strategies as mature.[7] Call it the "gen AI paradox": For all the energy, investment, and potential surrounding the technology, at-scale impact has yet to materialize for most organizations.

Exhibit 1

Gen AI has accelerated AI deployment overall.

Organizations that use AI in at least 1 business function,[1] % of respondents



[1] In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.
Source: McKinsey Global Surveys on AI

McKinsey & Company

At the heart of the gen AI paradox lies an imbalance between horizontal and vertical use cases

Many organizations have deployed horizontal use cases, such as enterprise-wide copilots and chatbots; nearly 70 percent of Fortune 500 companies, for example, use Microsoft 365 Copilot.[8] These tools are widely seen as levers to enhance individual productivity by helping employees save time on routine tasks and access and synthesize information more efficiently. But these improvements, while real, tend to be spread thinly across employees. As a result, they are not easily visible in terms of top- or bottom-line results.

[7] 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.
[8] Satya Nadella, "Microsoft Fiscal Year 2025 First Quarter Earnings Conference Call," Microsoft, October 30, 2024.

By contrast, vertical use cases—those embedded into specific business functions and processes—have seen limited scaling in most companies despite their higher potential for direct economic impact (Exhibit 2). Fewer than 10 percent of use cases deployed ever make it past the pilot stage, according to McKinsey research.[9] Even when they have been fully deployed, these use cases typically have supported only isolated steps of a business process and operated in a reactive mode when prompted by a human, rather than functioning proactively or autonomously. As a result, their impact on business performance also has been limited.

[9] *New at McKinsey Blog,* "McKinsey's ecosystem of strategic alliances brings the power of generative AI to clients," April 2, 2024.

Exhibit 2

Across business functions, gen AI use cases tend to fall into two categories: horizontal and vertical.

Example gen AI use cases across the enterprise and specific functions

	Functions					
	R&D	Procurement	Supply chain	Marketing and sales	Customer service	**Key features**
Horizontal use cases *Enterprise-wide*	**Employee copilots** Employee use of copilots, such as Microsoft 365 Copilot, Google Workspace AI, and Zoom AI Companion					Generates meeting minutes, writes emails, and translates text
	Chatbots Internal deployments of general-purpose tools, such as ChatGPT, Gemini, and Claude					Secured access to external GPT capabilities enriched with company-specific data
Vertical use cases *Function-specific*	**Idea generator** Summarizes research trends and proposes areas to explore	**Negotiation assistant** Suggests negotiation levers and next best actions	**Supply risk assessor** Generates risk profiles for suppliers to boost sourcing resilience	**Sales assistant** Prioritizes and qualifies accounts	**Support assistant** Supports staffers with suggested replies and knowledge base lookups	Function-specific copilots, triggered by a human prompt
	Research curator Auto-tags and organizes incoming scientific articles by topic	**Request for proposal (RPF) generator** Auto-generates complete RFPs	**Demand forecaster** Predicts demand based on internal and external data	**Insights provider** Provides campaign performance insights based on predefined KPIs and frameworks	**Ticket categorizer** Prioritizes tickets based on level of urgency	Individual task automation within a business workflow, triggered by a human prompt

McKinsey & Company

What accounts for this imbalance? For one thing, horizontally deployed copilots such as Microsoft Copilot or Google AI Workspace are accessible, off-the-shelf solutions that are relatively easy to implement. (In many cases, enabling Microsoft Copilot is as simple as activating an extension to an existing Office 365 contract, requiring no redesign of workflows or major change management efforts.) Rapid deployment of enterprise chatbots also has been driven by risk mitigation concerns. As employees began experimenting with external large language models (LLMs) such as ChatGPT, many organizations implemented internal, secure alternatives to limit data leakage and ensure compliance with corporate security policies.

The limited deployment and narrow scope of vertical use cases can in turn be attributed to six primary factors:

- *Fragmented initiatives.* At many companies, vertical use cases have been identified through a bottom-up, highly granular approach within individual functions. In fact, fewer than 30 percent of companies report that their CEOs sponsor their AI agenda directly.[10] This has led to a proliferation of disconnected micro-initiatives and a dispersion of AI investments, with limited coordination at the enterprise level.

- *Lack of mature, packaged solutions.* Unlike off-the-shelf horizontal applications, such as copilots, vertical use cases often require custom development. As a result, teams are frequently forced to build from scratch, using emerging, fast-evolving technologies they have limited experience with. While many companies have invested in data scientists to develop AI models, they often lack MLOps engineers, who are critical to industrialize, deploy, and maintain those models in production environments.

- *Technological limitations of LLMs.* Despite their impressive capabilities, the first generation of LLMs faced limitations that significantly constrained their deployment at enterprise scale. First, LLMs can produce inaccurate outputs, which makes them difficult to trust in environments where precision and repeatability are essential. What’s more, despite their power, LLMs are fundamentally passive; they do not act unless prompted and cannot independently drive workflows or make decisions without human initiation. LLMs also have struggled to handle complex workflows involving multiple steps, decision points, or branching logic. Finally, many current LLMs have limited persistent memory, making it difficult to track context over time or operate coherently across extended interactions.

- *Siloed AI teams.* AI centers of excellence have played a crucial role in accelerating awareness and experimentation across many organizations. However, in many cases, these teams have operated in silos—developing AI models independently from core IT, data, or business functions. This autonomy, while useful for rapid prototyping, has often made solutions difficult to scale because of poor integration with enterprise systems, fragmented data pipelines, or a lack of operational alignment.

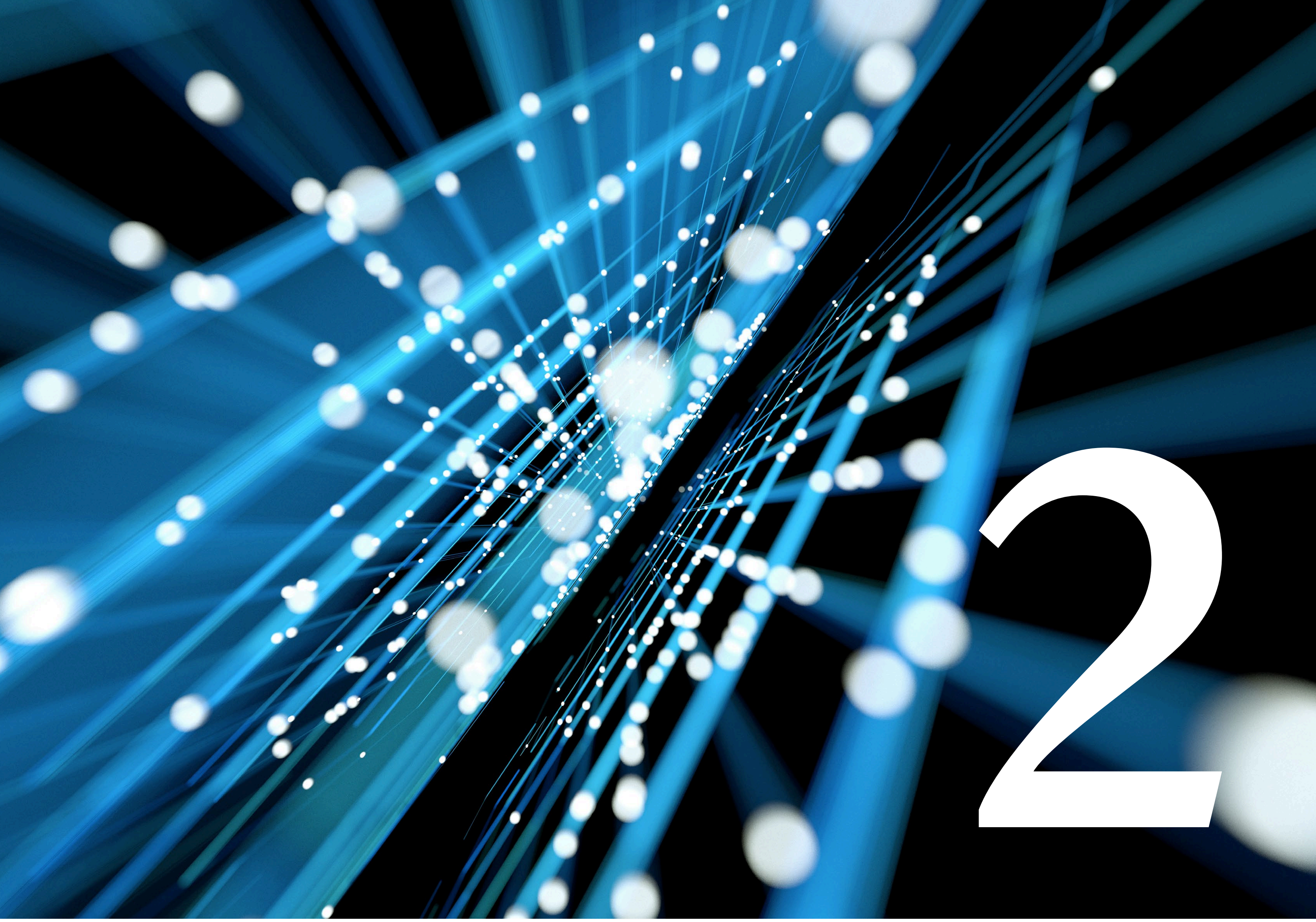
- *Data accessibility and quality gaps.* These gaps tend to exist for both structured and unstructured data, with unstructured material remaining largely ungoverned in most organizations.

- *Cultural apprehension and organizational inertia.* In many organizations, AI deployments have encountered implicit resistance from business teams and middle management due to fear of disruption, uncertainty around job impact, and lack of familiarity with the technology.

[10] “The state of AI: How organizations are rewiring to capture value,” McKinsey, March 12, 2025.

Despite its limited bottom-line impact so far, the first wave of gen AI has been far from wasted. It has enriched employee capabilities and enabled broad experimentation, accelerated AI familiarity across functions, and helped organizations build essential capabilities in prompt engineering, model evaluation, and governance. All of which has laid the groundwork for a more integrated and transformative second phase—the emerging age of AI agents.[11]

[11] Lareina Yee, Michael Chui, Roger Roberts, and Stephen Xu, "Why agents are the next frontier of generative AI," *McKinsey Quarterly,* July 24, 2024.



2

From paradox to payoff: How agents can scale AI

Key points:

- By automating complex business workflows, agents unlock the full potential of vertical use cases. Forward-looking companies are already harnessing the power of agents to transform core processes.
- To realize the potential of agents, companies must reinvent the way work gets done—changing task flows, redefining human roles, and building agent-centric processes from the ground up.
- Accomplishing this will require a new paradigm for AI architecture—the agentic AI mesh—capable of integrating both custom-built and off-the-shelf agents. But the bigger challenge will not be technical. It will be human: earning trust to drive adoption and establishing the proper governance protocols.

The breakthrough: Automating complex business workflows unlocks the full potential of vertical use cases

LLMs have revolutionized how organizations interact with data—enabling information synthesis, content generation, and natural language interaction. But despite their power, LLMs have been fundamentally reactive and isolated from enterprise systems, largely unable to retain memory of past interactions or context across sessions or queries. Their role has been largely limited to enhancing individual productivity through isolated tasks. AI agents mark a major evolution in enterprise AI—extending gen AI from reactive content generation to autonomous, goal-driven execution. Agents can understand goals, break them into subtasks, interact with both humans and systems, execute actions, and adapt in real time—all with minimal human intervention. They do so by combining LLMs with additional technology components providing memory, planning, orchestration, and integration capabilities.

With these new capabilities, AI agents expand the potential of horizontal solutions, upgrading general-purpose copilots from passive tools into proactive teammates that don't just respond to prompts but also monitor dashboards, trigger workflows, follow up on open actions, and deliver relevant insights in real time. But the real breakthrough comes in the vertical realm, where agentic AI enables the automation of complex business workflows involving multiple steps, actors, and systems—processes that were previously beyond the capabilities of first-generation gen AI tools.

Agents deliver more than efficiency—they supercharge operational agility and unlock new revenue opportunities

On the operations side, agents take on routine, data-heavy tasks so humans can focus on higher-value work. But they go further, transforming processes in five ways:

— *Agents accelerate execution by eliminating delays between tasks and by enabling parallel processing.* Unlike in traditional workflows that rely on sequential handoffs, agents can coordinate and execute multiple steps simultaneously, reducing cycle time and boosting responsiveness.

— *Agents bring adaptability.* By continuously ingesting data, agents can adjust process flows on the fly, reshuffling task sequences, reassigning priorities, or flagging anomalies before they cascade into failures. This makes workflows not only faster but smarter.

— *Agents enable personalization.* By tailoring interactions and decisions to individual customer profiles or behaviors, agents can adapt the process dynamically to maximize satisfaction and outcomes.

— *Agents bring elasticity to operations.* Because agents are digital, their execution capacity can expand or contract in real time depending on workload, business seasonality, or unexpected surges—something difficult to achieve with fixed human resource models.

— *Agents also make operations more resilient.* By monitoring disruptions, rerouting operations, and escalating only when needed, they keep processes running—whether it's supply chains navigating port delays or service workflows adapting to system outages.

In a complex supply chain environment, for example, an AI agent could act as an autonomous orchestration layer across sourcing, warehousing, and distribution operations. Connected to internal systems (such as the supply chain planning system or the warehouse management system) and external data sources (such as weather forecasts, supplier feeds, and demand signals), the agent could continuously forecast demand. It could then identify risks, such as delays or disruptions, and dynamically replan transport and inventory flows. Selecting the optimal transport mode based on cost, lead time, and environmental impact, the agent could reallocate stock across warehouses, negotiate directly with external systems, and escalate decisions requiring strategic input. The result: improved service levels, reduced logistics costs, and lower emissions.

Agents can also help spur top-line growth by amplifying existing revenue streams and unlocking entirely new ones:

— *Amplifying existing revenues.* In e-commerce, agents embedded into online stores or apps could proactively analyze user behavior, cart content, and context (for example, seasonality or purchase history) to surface real-time upselling and cross-selling offers. In finance, agents might help customers discover suitable financial products such as loans, insurance plans, or investment portfolios, providing tailored guidance based on financial profiles, life events, and user behavior.

— *Creating new revenue streams.* For industrial companies, agents embedded in connected products or equipment could monitor usage, detect performance thresholds, and autonomously unlock features or trigger maintenance actions—enabling pay-per-use, subscription, or performance-based models of creating revenue. Similarly, service organizations could encapsulate internal expertise—legal reasoning, tax interpretation, and procurement best practices—into AI agents offered as software-as-a-service tools or APIs to clients, partners, or smaller businesses lacking in-house expertise.

In short, agentic AI doesn't just automate. It redefines how organizations operate, adapt, and create value.

No longer science fiction: Forward-looking companies are harnessing the power of agents

The following case studies demonstrate how QuantumBlack helps organizations build agent workforces—with outcomes that extend far beyond efficiency gains.

Case study 1: How a bank used hybrid 'digital factories' for legacy app modernization

The problem: A large bank needed to modernize its legacy core system, which consisted of 400 pieces of software—a massive undertaking budgeted at more than $600 million. Large teams of coders tackled the project using manual, repetitive tasks, which resulted in difficulty coordinating across silos. They also relied on often slow, error-prone documentation and coding. While first-generation gen AI tools helped accelerate individual tasks, progress remained slow and laborious.

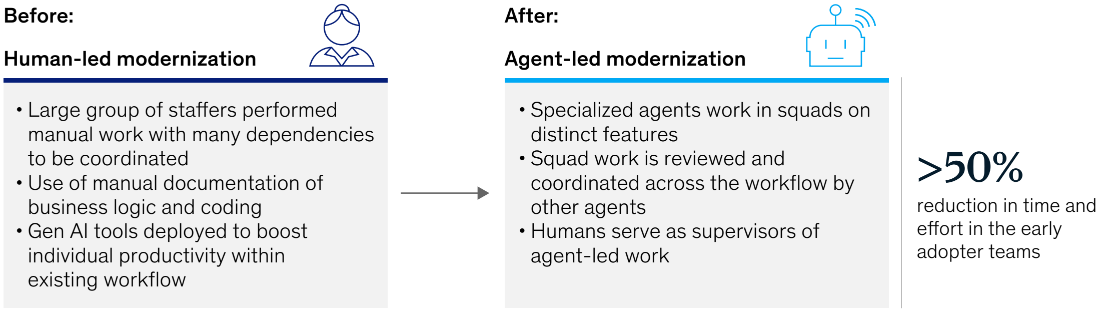
The agentic approach: Human workers were elevated to supervisory roles, overseeing squads of AI agents, each contributing to a shared objective in a defined sequence (Exhibit 3). These squads retroactively document the legacy application, write new code, review the code of other agents, and integrate code into features that are later tested by other agents prior to delivery of the end product. Freed from repetitive, manual tasks, human supervisors guide each stage of the process, enhancing the quality of deliverables and reducing the number of sprints required to implement new features.

Impact: More than 50 percent reduction in time and effort in the early adopter teams

Exhibit 3

A large bank upgraded its legacy tech stack with a hybrid AI–human digital factory.

Example: Banking modernization



Before:

Human-led modernization

- Large group of staffers performed manual work with many dependencies to be coordinated
- Use of manual documentation of business logic and coding
- Gen AI tools deployed to boost individual productivity within existing workflow

After:

Agent-led modernization

- Specialized agents work in squads on distinct features
- Squad work is reviewed and coordinated across the workflow by other agents
- Humans serve as supervisors of agent-led work

>50%
reduction in time and effort in the early adopter teams

McKinsey & Company

Case study 2: How a research firm boosted data quality to derive deeper market insights

The problem: A market research and intelligence firm was devoting substantial resources to ensure data quality, relying on a team of more than 500 people whose responsibilities included gathering data, structuring and codifying it, and generating tailored insights for clients. The process, conducted manually, was prone to error, with a staggering 80 percent of mistakes identified by the clients themselves.

The agentic approach: A multiagent solution autonomously identifies data anomalies and explains shifts in sales or market share. It analyzes internal signals, such as changes in product taxonomy, and external events identified via web searches, including product recalls or severe weather. The most influential drivers are synthesized, ranked, and prepared for decision-makers. With advanced search and contextual reasoning, the agents often surface insights that would be difficult for human analysts to uncover manually. While not yet in production, the system is fully functional and has demonstrated strong potential to free up analysts for more strategic work.

Impact: More than 60 percent potential productivity gain and expected savings of more than $3 million annually

Case study 3: How a bank reimagined the way it creates credit-risk memos
The problem: Relationship managers (RMs) at a retail bank were spending weeks writing and iterating credit-risk memos to help make credit decisions and fulfill regulatory requirements (Exhibit 4). This process required RMs to manually review and extract information from at least ten different data sources and develop complex nuanced reasoning across interdependent sections—for instance, loan, revenue, and cash joint evolution.

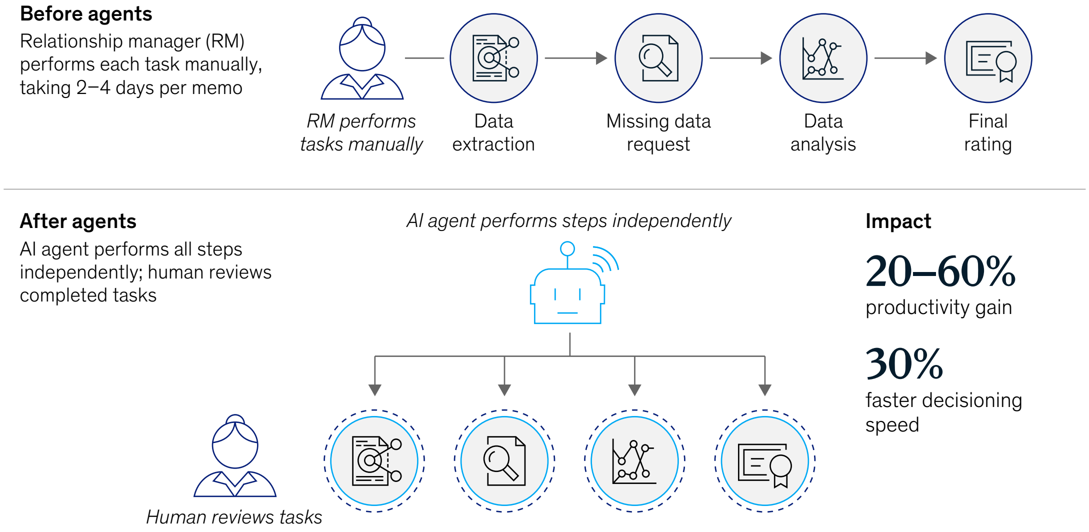
The agentic approach: In close collaboration with the bank's credit-risk experts and RMs, a proof of concept was developed to transform the credit memo workflow using AI agents. The agents assist RMs by extracting data, drafting memo sections, generating confidence scores to prioritize review, and suggesting relevant follow-up questions. In this model, the analyst's role shifts from manual drafting to strategic oversight and exception handling.

Impact: A potential 20 to 60 percent increase in productivity, including a 30 percent improvement in credit turnaround

Exhibit 4

A retail bank used AI agents to reinvent the process of creating credit-risk memos.

Example: Retail bank process



Before agents

Relationship manager (RM) performs each task manually, taking 2–4 days per memo

After agents

AI agent performs all steps independently; human reviews completed tasks

Impact

20–60%
productivity gain

30%
faster decisioning speed

McKinsey & Company

Maximizing value from AI agents requires process reinvention

Realizing AI's full potential in the vertical realm requires more than simply inserting agents into legacy workflows. It instead calls for a shift in design mindset—from automating tasks within an existing process to reinventing the entire process with human and agentic coworkers. That's because when agents are embedded into a legacy process without redesign, they typically serve as faster assistants—generating content, retrieving data, or executing predefined steps. But the process itself remains sequential, rule bound, and shaped by human constraints.

Reinventing a process around agents means more than layering automation on top of existing workflows—it involves rearchitecting the entire task flow from the ground up. That includes reordering steps, reallocating responsibilities between humans and agents, and designing the process to fully exploit the strengths of agentic AI: parallel execution that collapses cycle time, real-time adaptability that reacts to changing conditions, deep personalization at scale, and elastic capacity that flexes instantly with demand.

Consider a hypothetical customer call center. Before introducing AI agents, the facility was using gen AI tools to assist human support staff by retrieving articles from knowledge bases, summarizing ticket histories, and helping draft responses. While this assistance improved speed and reduced cognitive load, the process itself remained entirely manual and reactive, with human agents still managing every step of diagnosis, coordination, and resolution. The productivity improvement potential was modest, typically boosting resolution time and productivity between 5 and 10 percent.

Now imagine that the call center introduces AI agents but largely preserves the existing workflow—agents are added to assist at specific steps without reconfiguring how work is routed, tracked, or resolved end-to-end. Agents can classify tickets, suggest likely root causes, propose resolution paths, and even autonomously resolve frequent, low-complexity issues (such as password resets). While the impact here can be increased—an estimated 20 to 40 percent savings in time and a 30 to 50 percent reduction in backlog—coordination friction and limited adaptability prevent true breakthrough gains.

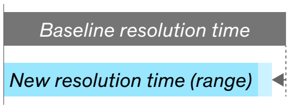
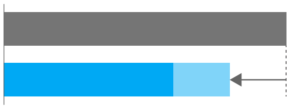
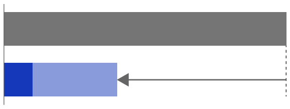
But the real shift occurs at the third level, when the call center's process is reimagined around agent autonomy. In this model, AI agents don't just respond—they proactively detect common customer issues (such as delayed shipments, failed payments, or service outages) by monitoring patterns across channels, anticipate likely needs, initiate resolution steps automatically (such as issuing refunds, reordering items, or updating account details), and communicate directly with customers via chat or email. Human agents are repositioned as escalation managers and service quality overseers, who are brought in only when agents detect uncertainty or exceptions to typical patterns. Impact at this level is transformative. This could allow a radical improvement of customer service desk productivity. Up to 80 percent of common incidents could be resolved autonomously, with a reduction in time to resolution of 60 to 90 percent (Exhibit 5).

Of course, not every business process requires full reinvention. Simple task automation is sufficient for highly standard, repetitive workflows with limited variability—such as payroll processing, travel expense approvals, or password resets—where gains come primarily from reducing manual effort. In contrast, processes that are complex, cross-functional, prone to exceptions, or tightly linked to business performance often warrant full redesign. Key indicators that call for reinvention include high coordination overhead, rigid sequences that delay responsiveness, frequent human intervention for decisions that could be data driven, and opportunities for dynamic adaptation or personalization. In these cases, redesigning the process around the agent's ability to orchestrate, adapt, and learn delivers far greater value than simply speeding up existing workflows.

Exhibit 5

Agents hold the key to breaking through—if processes are reinvented, not just optimized.

Example: Call center performance comparison

	Gen AI–enabled	Agent-enabled (optimized)	Agent-enabled (reinvented)
Description	Human manages every step of the workflow, assisted by gen AI tools to help retrieve knowledge-based articles, summarize ticket history, and draft responses	AI agent automates discrete tasks within existing workflow, such as ticket classification, suggestion of likely root causes, and resolution of frequent, low-complexity issues	Process is redesigned around agent autonomy: AI agents proactively detect incidents, diagnose issues, and initiate resolutions automatically
Estimated impact	 **5–10%** average reduction in resolution time	 **20–40%** average reduction in resolution time	 **60–90%** average reduction in resolution time **80%** of level 1 incidents resolved automatically

McKinsey & Company

A new AI architecture paradigm—the agentic AI mesh—is required to orchestrate value in the agentic era

To scale agents, companies will need to overcome a threefold challenge: handling the newfound risks that AI agents bring, blending custom and off-the-shelf agentic systems, and staying agile amid fast-evolving tech (while avoiding lock-ins).

— *Managing a new wave of risks.* Agents introduce a new class of systemic risks that traditional gen AI architectures, designed primarily for isolated LLM-centric use cases, were never built to handle: uncontrolled autonomy, fragmented system access, lack of observability and traceability, expanding surface of attack, and agent sprawl and duplication. What starts as intelligent automation can quickly become operational chaos—unless it is built on a foundation that prioritizes control, scalability, and trust.

— *Blending custom and off-the-shelf agents.* To fully capture the transformative potential of AI agents, organizations must go beyond simply activating agents embedded in software suites. These off-the-shelf agents may streamline routine workflows, but they rarely unlock strategic advantage. Realizing the full potential of agentic AI will require the development of custom-built agents for high-impact processes, such as end-to-end customer resolution, adaptive supply chain orchestration, or complex decision-making. These agents must be deeply aligned with the company's logic, data flows, and value creation levers—making them difficult to replicate and uniquely powerful.

— *Staying agile amid fast-evolving tech.* Agentic AI is a new technology area, and solutions are evolving very rapidly. Agents will have to support workflows across multiple systems and should not be hardwired within a specific platform. An evolutive and vendor-agnostic architecture is therefore needed.

These challenges cannot be addressed by merely bolting new components, such as memory stores or orchestration engines, on top of existing gen AI stacks. While such capabilities are necessary, they are not sufficient. What's needed is a fundamental architectural shift: from static, LLM-centric infrastructure to a dynamic, modular, and governed environment built specifically for agent-based intelligence—the agentic AI mesh.

The agentic AI mesh is a composable, distributed, and vendor-agnostic architectural paradigm that enables multiple agents to reason, collaborate, and act autonomously across a wide array of systems, tools, and language models—securely, at scale, and built to evolve with the technology. At the heart of this paradigm are five mutually reinforcing design principles:

- *Composability.* Any agent, tool, or LLM can be plugged into the mesh without system rework.

- *Distributed intelligence.* Tasks can be decomposed and resolved by networks of cooperating agents.

- *Layered decoupling.* Logic, memory, orchestration, and interface functions are decoupled to maximize modularity.

- *Vendor neutrality.* All components can be independently updated or replaced as technology advances, avoiding vendor lock-in and future-proofing the architecture. In particular, open standards such as the Model Context Protocol (MCP) and Agent2Agent (A2A) are preferred to proprietary protocols.

- *Governed autonomy.* Agent behavior is proactively controlled via embedded policies, permissions, and escalation mechanisms that ensure safe, transparent operation.

The agentic AI mesh acts as the connective and orchestration layer that enables large-scale, intelligent agent ecosystems to operate safely and efficiently, and continuously evolve. It allows companies to coordinate custom-built and off-the-shelf agents within a unified framework, support multiagent collaboration by allowing agents to share context and delegate tasks, and mitigate key risks such as agent sprawl, autonomy drift, and lack of observability—all while preserving the agility required for a rapid technology evolution (see sidebar "Seven interconnected capabilities of the AI agentic mesh").

Beyond this architectural evolution, organizations will also have to revisit their LLM strategies. At the core of every custom agent lies a foundation model—the reasoning engine that powers perception, decision-making, and interaction. In the agentic era, the requirements placed on LLMs evolve significantly. Agents are not passive copilots—they are autonomous, persistent, embedded systems. This creates five critical categories of LLM requirements, each aligned with specific deployment contexts, for which different kinds of models will be relevant (see sidebar "Foundational models for agents: Five new requirements").

Finally, to truly scale agent deployment across the enterprise, the enterprise systems themselves must also evolve.

In the short term, APIs—protocols that allow different software applications to communicate and exchange data—will remain the primary interface for agents to interact with enterprise systems. But in the long term, APIs alone will not suffice. Organizations must begin reimagining their IT architectures around an agent-first model—one in which user interfaces, logic, and data access layers are natively designed for machine interaction rather than human navigation. In such a model, systems are no longer organized around screens and forms but around machine-readable interfaces, autonomous workflows, and agent-led decision flows.

Seven interconnected capabilities of the AI agentic mesh

The emerging architecture for agentic AI relies on seven interconnected capabilities:

1. *Agent and workflow discovery* maintains a dynamic catalog of all organizational agents and workflows, enabling reuse across teams and enforcing policies on agent use.

2. *AI asset registry* centralizes governance of system prompts, agent instructions, large-language-model (LLM) configurations, tool definitions, and golden records while creating policies about version control and access.

3. *Observability* provides end-to-end tracing of workflows spanning agentic and procedural systems through standardized metrics, audit logs, and diagnostic capabilities.

4. *Authentication and authorization* enforce fine-grain access controls for communication among agentic systems, procedural systems, and LLMs, enforcing security policies and limiting the "blast radius" of compromised systems or agents.

5. *Evaluations* deliver comprehensive testing of agent pipelines to ensure accuracy and compliance over time.

6. *Feedback management* enables continuous improvement through automated feedback loops that capture performance metrics to evolve agent configurations.

7. *Compliance and risk management* embed policy controls, compliance agents, and ethical guardrails to ensure workflows meet regulatory and institutional standards.

This shift is already underway. Microsoft is embedding agents into the core of Dynamics 365 and Microsoft 365 via Copilot Studio; Salesforce is expanding Agentforce into a multiagent orchestration layer; SAP is rearchitecting its Business Technology Platform (BTP) to support agent integration through Joule. These changes signal a broader transition: The future of enterprise software is not just AI-augmented—it is agent-native.

The main challenge won't be technical—it will be human

As agents evolve from passive copilots to proactive actors—and scale across the enterprise—the complexity they introduce will be not only technical but mostly organizational. The real challenge lies in coordination, judgment, and trust. This organizational complexity will play out most visibly across three dimensions: how humans and agents cohabit day-to-day workflows; how organizations establish governance over systems that can act autonomously; and how they prevent unchecked sprawl as agent creation becomes increasingly democratized.

— *Human–agent cohabitation.* Agents won't just assist humans—they'll act alongside them. This raises nuanced questions about interaction and coexistence: When should an agent take initiative? When should it defer? How do we maintain human agency and oversight without slowing down the very

Foundational models for agents: Five new requirements

For LLMs to function properly in the agentic age, they will need to evolve in a number of critical ways:

1. *Low-latency inference for real-time responsiveness.* Agents embedded in workflows (such as service operations or IT alerts) require subsecond response times with predictable latency, even under compute constraints. Illustrative examples of relevant models include Mistral Small (Mistral AI), Llama 3 8B (Meta), Gemini Nano (Google), and Claude Haiku (Anthropic).

2. *Fine-tuning and controllability for domain-specific agents.* Agents operating in regulated or knowledge-intensive domains (such as finance, legal, and healthcare) need large language models (LLMs) that can be fine-tuned, grounded in enterprise knowledge, and instrumented with external tools (such as RAG and APIs). Illustrative examples of relevant models are Mistral Small and Mistral 8x7B (open weight and fine-tunable, Mistral AI), and Llama 3 8B and 70B (fine-tunable, Meta).

3. *Lightweight deployment for embedded and edge agents.* In cases such as the Internet of Things, field devices, or privacy-sensitive environments, agents must be embedded directly into software or hardware, with minimal compute and memory footprint. Illustrative examples of relevant models include Mistral Small (Mistral AI), Gemini Nano (Google), Llama 3 8B (Meta), and Phi-2 (Microsoft).

4. *Scalable multiagent orchestration across the enterprise.* Enterprises deploying hundreds or thousands of agents require LLMs that can scale efficiently and cost-effectively, ideally using sparse architectures or a mixture of experts. Illustrative examples of relevant models include Mixtral (Mistral AI), Grok-1 (xAI), GPT-3.5 Turbo (OpenAI), and Command R+ (Cohere).

5. *Sovereignty, auditability, and geopolitical resilience for autonomous agents.* Agents embedded in core operations—particularly in public, financial, and critical-infrastructure sectors—must ensure compliance, data sovereignty, traceability, and geopolitical autonomy. This includes avoiding reliance on APIs that are hosted abroad, ensuring data residency, and resisting extraterritorial legal exposure (for example, OpenAI or Anthropic subject to US subpoenas). Illustrative examples of relevant models include Mistral Small/Mixtral (Mistral AI), Falcon 180B (TII UAE), and BloomZ/Bloom (BigScience).

benefits agents bring? Building clarity around these roles will take time, experimentation, and cultural adjustment. Trust won’t come from technical performance alone—it will hinge on how transparently agents communicate, how predictably they behave, and how intuitively they integrate into daily workflows.

— *Autonomy control.* What makes agents powerful—their ability to act independently—also introduces ambiguity. Unlike traditional tools, agents don’t wait to be instructed. They respond, adapt, and sometimes surprise. Navigating this new reality means confronting edge cases: What if an agent executes too aggressively? Or fails to escalate a subtle issue? The challenge is not to eliminate autonomy but to make it intelligible and aligned with organizational expectations. That alignment won’t be static.

It will need to evolve as agents learn, systems shift, and trust deepens. Control mechanisms must also address the risk of hallucinations, or plausible but inaccurate outputs agents may produce.

- *Sprawl containment.* As in the early days of robotic process automation, there's a real risk of agent sprawl—the uncontrolled proliferation of redundant, fragmented, and ungoverned agents across teams and functions. As low-code and no-code platforms make agent creation accessible to anyone, organizations risk a new kind of shadow IT: agents that multiply across teams, duplicate efforts, or operate without oversight. How do we avoid fragmentation? Who decides what gets built—and what gets retired? Without structured governance, design standards, and life cycle management, agent ecosystems can quickly become fragile, redundant, and unscalable.

Agents unlock the full potential of vertical use cases, offering companies a path to generate value well beyond efficiency gains. But realizing that potential requires a reimagined approach to AI transformation—one tailored to the unique nature of agents and capable of addressing the lingering limitations they alone cannot resolve. This approach is the subject of our next chapter.

3

AI transformation at a tipping point: The CEO mandate in the agentic era

Key points:

- Generating impact in the agentic era requires organizations to shift from scattered initiatives to strategic programs; from use cases to business processes; from siloed AI teams to cross-functional transformation squads; and from experimentation to industrialized, scalable delivery.

- To scale agents, organizations will also need to set a new foundation by upskilling the workforce, adapting the technology infrastructure, and developing new governance structures for agents.

- The time has come to bring the gen AI experimentation phase to an end—a pivot only the CEO can make.

Scaling impact in the agentic era requires a reset of the AI transformation approach

Unlike gen AI tools that could be easily plugged into existing workflows, AI agents demand a more foundational shift, one that requires rethinking business processes and enabling deep integration with enterprise systems. McKinsey has a proven Rewired playbook for digital transformations.[12] To capitalize on the agentic opportunity, organizations must build on that, fundamentally reshaping their AI transformation approach across four dimensions:

— *Strategy: From scattered tactical initiatives to strategic programs.* With agentic AI set to reshape the foundations of competition, organizations must move beyond bottom-up use case identification and directly align AI initiatives with their most critical strategic priorities. This means not only translating existing goals—such as enhancing operational efficiency, improving customer intimacy, or strengthening compliance—into AI-addressable transformation domains, but also adopting a forward-looking lens. Executives must challenge their organizations to look beyond the boundaries of today's operating model and explore how AI can be used to reimagine entire segments of the business, create new revenue streams, and build competitive moats that will define leadership in the next decade.

— *Unit of transformation: From use case to business processes.* In the early wave of gen AI adoption, most vertical initiatives focused on plugging a solution into a specific step of an existing process—which tended to deliver narrow gains without changing the overall structure of how work is done. With AI agents, the paradigm shifts entirely. Opportunity now lies not in optimizing isolated tasks but in transforming entire business processes by embedding agents throughout the value chain. As a result, AI initiatives should no longer be scoped around a single use case, but instead around the end-to-end reinvention of a full process or persona journey. In vertical domains, this means moving from the question, "Where can I use AI in this function?" to "What would this function look like if agents ran 60 percent of it?" It involves rethinking workflows, decision logic, human–system interactions, and performance metrics across the board.

[12] Eric Lamarre, Kate Smaje, and Rodney Zemmel, "Rewired to outcompete," *McKinsey Quarterly,* June 20, 2023.

— *Delivery model: From siloed AI teams to cross-functional transformation squads.* AI centers of excellence have played a key role in accelerating AI awareness and experimentation across organizations. However, this model reaches its limits in the agentic era—in which agents are deeply embedded into enterprise systems, operate across complex business processes, and rely on high-quality data as their primary fuel. In this context, AI initiatives can no longer be delivered by isolated, specialized AI teams. To succeed at scale, organizations must shift to a cross-functional delivery model, anchored in durable transformation squads composed of business domain experts, process designers, AI and MLOps engineers, IT architects, software engineers, and data engineers.

— *Implementation process: From experimentation to industrialized, scalable delivery.* While the previous phase rightly focused on exploring the potential of gen AI, organizations must now shift to an industrialized delivery model, in which solutions are designed from the outset to scale, both technically and financially. This requires organizations to anticipate the full set of technical prerequisites for enterprise deployment—notably in terms of system integration, day-to-day monitoring, and release management, but also to rigorously estimate future running costs and design a solution to minimize them. Unlike traditional IT systems—for which annual run costs typically represent 10 to 20 percent of initial build costs[13]—gen AI solutions, especially at scale, can incur recurring costs that exceed the initial build investment. Designing for scalability must therefore include not just technical robustness but also economic sustainability, especially for high-volume applications.

Four critical enablers are required to effectively operate in the agentic era

Redesigning the approach to AI transformation is an important step, but it is not enough. To unlock their full potential at scale, organizations must also activate a robust set of enablers that support the structural, cultural, and technical shifts required to integrate agents into day-to-day operations. These enablers span four dimensions—people, governance, technology architecture, and data—each of which is a foundation for scalable, secure, and high-impact deployment of agents across the enterprise.

— *People: Equip the workforce and introduce new roles.* The workforce must be equipped for new ways of working driven by human–agent collaboration. This involves fostering a “human + agent” mindset through cultural change, targeted training, and supporting early adopters as internal champions. New roles must also be introduced, such as prompt engineers to refine interactions, agent orchestrators to manage agent workflows, and human-in-the-loop designers to handle exceptions and build trust.

— *Governance: Ensure autonomy control and prevent agent sprawl.* With the rise of autonomous agents comes the need for strong governance to avoid risk and uncontrolled sprawl. Enterprises should define governance frameworks that establish agent autonomy levels, decision boundaries, behavior monitoring, and audit mechanisms. Policies for development, deployment, and usage must also be formalized, along with classification systems that group agents by function (such as task automators, domain orchestrators, and virtual collaborators), each with an appropriate oversight model.

— *Technology architecture: Build a foundation for interoperability and scale.* Agents, whether custom-built or off-the-shelf, must operate across a fragmented ecosystem of systems, data, and workflows. In the short term, organizations must evolve their AI architecture from LLM-centric setups to an agentic

[13] Aykut Atali, Chandra Gnanasambandam, and Bhargs Srivathsan, “Transforming infrastructure operations for a hybrid-cloud world,” McKinsey, October 9, 2019.

AI mesh. Beyond this first step, organizations should start preparing for their next-generation architecture, in which all enterprise systems will be reshuffled around agents in terms of user interface, business logic, and day-to-day operations.

— *Data: Accelerate data productization and address quality gaps in unstructured data.* Finally, agents depend on the quality and accessibility of enterprise data. Organizations must transition from use-case-specific data pipelines to reusable data products and extend data governance to unstructured data.

CEOs have a leadership challenge: Bringing the gen AI experimentation phase to a close

The rise of AI agents is more than just a technological shift. Agents represent a strategic inflection point that will redefine how companies operate, compete, and create value. To navigate this transition successfully, organizations must move beyond experimentation and pilot programs and enter a new phase of scaled, enterprise-wide transformation.

This pivot cannot be delegated—it must be initiated and led by the CEO. It will rely on three key actions:

— *Action 1: Conclude the experimentation phase and realign AI priorities.* Conduct a structured review to capture lessons learned, retire unscalable pilots, and formally close the exploratory phase. Refocus efforts on strategic AI programs targeting high-impact domains and processes.

— *Action 2: Redesign the AI governance and operating model.* Set up a strategic AI council involving business leaders, the chief human resources officer, the chief data officer, and the chief information officer. This council should oversee AI direction-setting; coordinate AI, IT, and data investments; and implement rigorous value-tracking mechanisms based on KPIs tied to business outcomes.

— *Action 3: Launch a first lighthouse transformation project and simultaneously initialize the agentic AI tech foundation.* Kick off a select number of high-impact agentic AI–driven workflow transformations in core business areas. In parallel, lay the groundwork for an agentic AI technology foundation by investing in key enablers—technology infrastructure, data quality, governance frameworks, and workforce readiness.

Conclusion

Like any truly disruptive technology, AI agents have the power to reshuffle the deck. Done right, they offer laggards a leapfrog opportunity to rewire their competitiveness. Done wrong—or not at all—they risk accelerating the decline of today's market leaders. This is a moment of strategic divergence.

While the technology will continue to evolve, it is already mature enough to drive real, transformative change across industries. But to realize the full promise of agentic AI, CEOs must rethink their approach to AI transformation—not as a series of scattered pilots but as focused, end-to-end reinvention efforts. That means identifying a few business domains with the highest potential and pulling every lever: from reimagining workflows to redistributing tasks between humans and machines to rewiring the organization based on new operating models.

Some leaders are already moving—not just by deploying fleets of agents but by rewiring their organizations to harness their full disruptive potential. (Moderna, for example, merged its HR and IT leadership[14]—signaling that AI is not just a technical tool but a workforce-shaping force.) This is a structural move toward a new kind of enterprise. Agentic AI is not an incremental step—it is the foundation of the next-generation operating model. CEOs who act now won't just gain a performance edge. They will redefine how their organizations think, decide, and execute.

The time for exploration is ending. The time for transformation is now.

Alexander Sukharevsky is a senior partner in McKinsey's London office, where **Dave Kerr** is a partner; **Klemens Hjartar** is a senior partner in the Copenhagen office; **Lari Hämäläinen** is a senior partner in the Seattle office; **Stéphane Bout** is a senior partner in the Lyon office; **Vito Di Leo** is a partner in the Zurich office; and **Guillaume Dagorret** is a senior fellow with the McKinsey Global Institute and is based in the Paris office.

The authors wish to thank Alena Fedorenko, Annie David, Clarisse Magnin, Lareina Yee, Larry Kanter, Michael Chui, Roger Roberts, Sarah Mulligan, Thomas Vlot, and Timo Mauerhoefer for their contributions to this report.

[14] Julien Dupont-Calbo, "L'IA n'est plus un outil, c'est un collègue": Moderna fusionne sa DRH et sa DSI, ["AI is no longer a tool, it's a colleague": Moderna merges its HR and IT departments], *Les Echos,* May 15, 2025.

McKinsey & Company
June 2025

Designed by Darby

www.mckinsey.com

@McKinsey
@McKinsey