McKinsey & Company | QuantumBlack AI by McKinsey

Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom

February 2024

McKinsey & Company | QuantumBlack AI by McKinsey

Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom

February 2024

Contents

Introduction: The promise and the challenge of generative AI

The emergence of generative AI (gen AI) presents both a challenge and a significant opportunity for leaders looking to steer their organizations into the future. How big is the opportunity? McKinsey research estimates that gen AI could add to the economy between $2.6 trillion and $4.4 trillion annually while increasing the impact of all artificial intelligence by 15 to 40 percent. In the technology, media, and telecommunications (TMT) space, new gen AI use cases are expected to unleash between $380 billion and $690 billion in impact—$60 billion to $100 billion in telecommunications, $80 billion to $130 billion in media, and about $240 billion to $460 billion in high tech. In fact, it seems possible that within the next three years, anything not connected to AI will be considered obsolete or ineffective.

Some leaders are moving to seize the moment and implement gen AI in their organizations at scale, but others remain in the pilot stage, and some have yet to decide what to do. If companies are to remain competitive and relevant in the coming years, it is essential that executives understand the potential impact of gen AI and develop the strategies necessary to incorporate it into their operations. Such strategies would involve an AI-native transformation, focused on building and managing the adoption of gen AI. McKinsey has conducted extensive research into how to embed gen AI to ensure that the technology delivers meaningful value. We've also spent much of the past year working with clients to create and then implement gen AI road maps. That combination of research and hands-on experience has allowed us to identify more than 100 gen AI use cases in TMT across seven business domains.[1]

Our experience working with clients already indicates the potential for telcos to achieve significant impact with gen AI across all key functions. The largest share of total impact will likely be in customer care and sales, which together would account for approximately 70 percent of total impact; network operations, IT, and support functions would round out the rest. The technology already is showing meaningful impact in enhancing interactions between employees and customers: the personalization of products and campaigns, improvements in sales effectiveness, and a reduction in time to market can spark a potential revenue increase of 3 to 5 percent. Customer care interactions—where as much as 50 percent of activity could be automated—have potential for a 30 to 45 percent increase in productivity while improving the customer experience and customer satisfaction scores. On the labor side, up to 70 percent of repetitive work activities could be automated via gen AI to improve productivity. There is also potential for new efficiencies in knowledge search, validation, and synthesis, where some 60 percent of activity has the potential for automation. And gen AI tools could boost developer productivity by 20 to 45 percent.

These areas provide rich soil for use cases. More challenging will be to go from sketching a road map to building proofs of concept to scaling successfully and capturing impact. Years of experience in designing and implementing digital transformations have taught us a lot, but gen AI's nature and speed of disruption are creating a new layer of uncertainty.

Becoming an AI-native organization at scale involves making the most of technology, data, and governance. Success follows when leaders embrace an operating model that leverages the strengths of both humans and machines; is rooted in agility, flexibility, and continuous learning; and is supported by strong data and analytics talent. Another condition of success is to invest in data quality and quantity, focusing on the data life cycle to ensure high-quality information for training the gen AI model. Building capabilities into the data architecture, such as vector databases and data pre- and post-processing pipelines, will enable the development of use cases. Talent, data, technology, governance—none of these can be an afterthought.

[1] Marketing and digital, sales and channels, customer care, customer strategy, support, additional areas, and new businesses.

Successful implementations share a clear vision and decisive approach. We advise that financial plans maintain or increase gen AI budgets over the next year. These budgets should include resources dedicated to gen AI for the shaping and crafting of bespoke solutions (for example, training large language models with telco-specific data, rather than implementing off-the-shelf ones) or partnerships with IT vendors to accelerate the timeline for implementation.

The AI journey has been shown to contain many challenges and learning opportunities, such as preparing and shifting an organization's culture, finding data sets of significant size, and addressing the interpretability of the outputs provided by models. Leaders should expect such daunting challenges as a shortage of talent, lack of organizational commitment and prioritization (including among C-level executives), and difficulties in justifying ROI for certain business cases, all amid a changing regulatory and ethics landscape that creates further uncertainty. But daunting does not have to mean impossible. Developing a system of protocols and guardrails (such as building "moderation" models to check outputs for different risks and ensure users receive consistent responses) will be a crucial step toward mitigating the new risks introduced by gen AI. Another key will be change management—involving end users in the model development process and deeply embedding technology into their operations.

This collection presents McKinsey's top insights on gen AI, providing a detailed examination of this technology's transformative potential for organizations. It offers top management guidance on how to prepare for the implementation of gen AI and explores the implications of gen AI's use by the TMT industries, especially telecommunications. The collection covers the essential requirements for deploying gen AI, including organizational readiness, data management, and technological considerations. It also emphasizes the importance of effectively managing risks associated with gen AI implementation. Furthermore, this compilation offers an overview of the future developments and advancements expected in the field of generative AI.

Gen AI will continue to evolve. New capabilities, such as the ability to analyze and comprehend images or audio, and an expanding ecosystem with marketplaces for GPT (generative pretrained transformers), are constantly emerging. For leaders, the stakes are high. But so are the opportunities. The next move from TMT players will define how they move from isolated cases to implementations at scale, from hype to impact.

Alex Singla
Senior Partner
Managing Partner
QuantumBlack
AI by McKinsey

Alexander Sukharevsky
Senior Partner
Managing Partner
QuantumBlack
AI by McKinsey

Brendan Gaffey
Senior Partner
Global Leader
TMT Practice

Noshir Kaka
Senior Partner
Global Leader
TMT Practice

Peter Dahlström
Senior Partner
Europe Leader
TMT Practice

Andrea Travasoni
Senior Partner
Global Leader
Telecom Operators
TMT Practice

Venkat Atluri
Senior Partner
Global Leader
Telecom Operators
TMT Practice

Tomás Lajous
Senior Partner
AI and Gen AI Leader
TMT Practice

Benjamim Vieira
Senior Partner
Digital and Analytics Leader
TMT Practice

Víctor García de la Torre
Associate Partner
TMT Practice

1

State of the art

McKinsey & Company

The economic potential of generative AI

The next productivity frontier

June 2023

Authors

Michael Chui
Eric Hazan
Roger Roberts
Alex Singla
Kate Smaje
Alexander Sukharevsky
Lareina Yee
Rodney Zemmel

1

Generative AI as a technology catalyst

To grasp what lies ahead requires an understanding of the breakthroughs that have enabled the rise of generative AI, which were decades in the making. ChatGPT, GitHub Copilot, Stable Diffusion, and other generative AI tools that have captured current public attention are the result of significant levels of investment in recent years that have helped advance machine learning and deep learning. This investment undergirds the AI applications embedded in many of the products and services we use every day.

But because AI has permeated our lives incrementally—through everything from the tech powering our smartphones to autonomous-driving features on cars to the tools retailers use to surprise and delight consumers—its progress was almost imperceptible. Clear milestones, such as when AlphaGo, an AI-based program developed by DeepMind, defeated a world champion Go player in 2016, were celebrated but then quickly faded from the public's consciousness.

ChatGPT and its competitors have captured the imagination of people around the world in a way AlphaGo did not, thanks to their broad utility—almost anyone can use them to communicate and create—and preternatural ability to have a conversation with a user. The latest generative AI applications can perform a range of routine tasks, such as the reorganization and classification

This article is excerpted from the full McKinsey report, The economic potential of generative AI: The next productivity frontier. *To read the full report, including details about the research, appendix, and acknowledgements, visit mck.co/genai.*

of data. But it is their ability to write text, compose music, and create digital art that has garnered headlines and persuaded consumers and households to experiment on their own. As a result, a broader set of stakeholders are grappling with generative AI's impact on business and society but without much context to help them make sense of it.

How did we get here? Gradually, then all of a sudden

For the purposes of this report, we define generative AI as applications typically built using foundation models. These models contain expansive artificial neural networks inspired by the billions of neurons connected in the human brain. Foundation models are part of what is called deep learning, a term that alludes to the many deep layers within neural networks. Deep learning has powered many of the recent advances in AI, but the foundation models powering generative AI applications are a step change evolution within deep learning. Unlike previous deep learning models, they can process extremely large and varied sets of unstructured data and perform more than one task.

Foundation models have enabled new capabilities and vastly improved existing ones across a broad range of modalities, including images, video, audio, and computer code. AI trained on these models can perform several functions; it can classify, edit, summarize, answer questions, and draft new content, among other tasks.

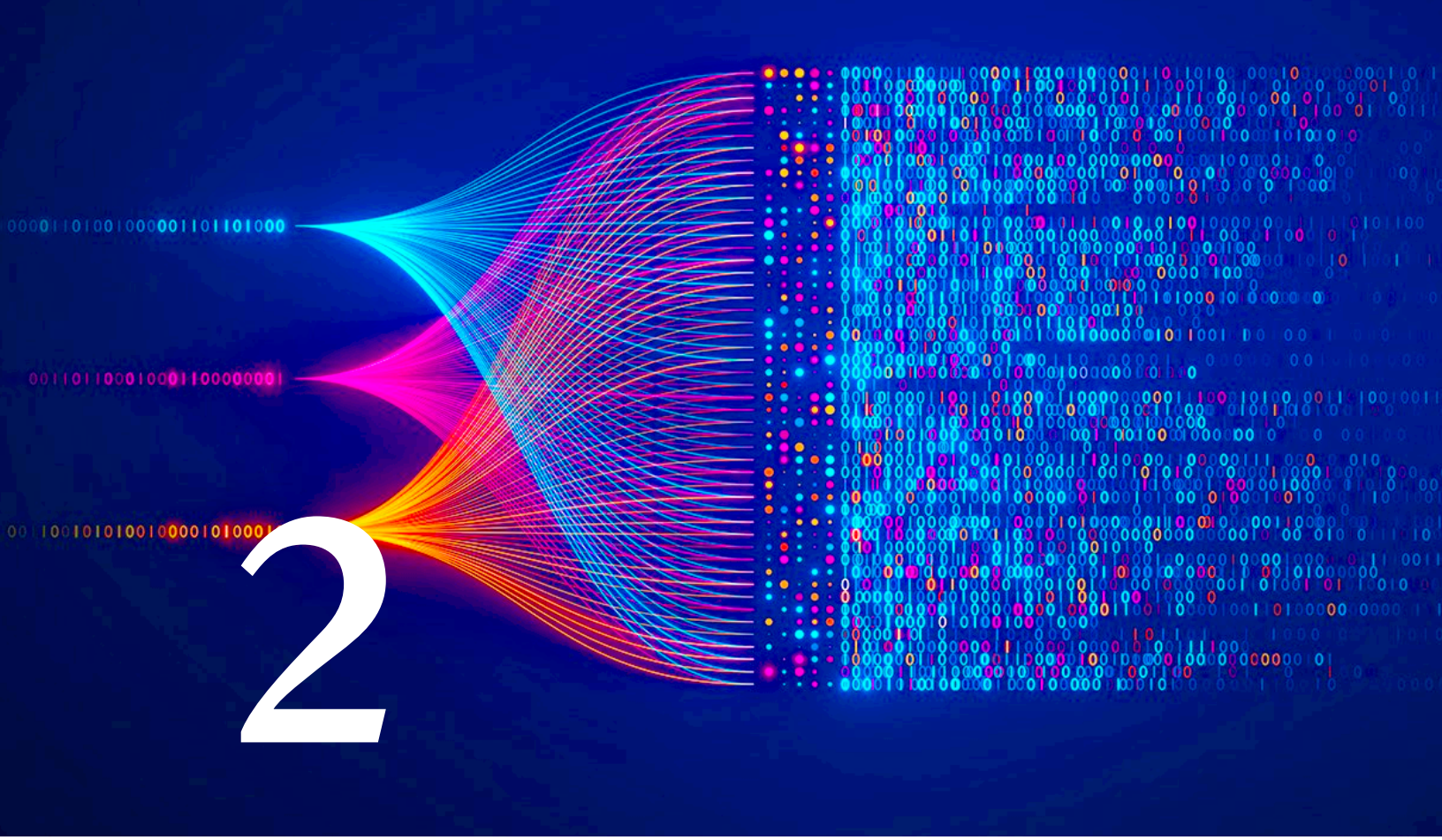
Continued innovation will also bring new challenges. For example, the computational power required to train generative AI with hundreds of billions of parameters threatens to become a bottleneck in development.[1] Further, there's a significant move—spearheaded by the open-source community and spreading to the leaders of generative AI companies themselves—to make AI more responsible, which could increase its costs.

Nonetheless, funding for generative AI, though still a fraction of total investments in artificial intelligence, is significant and growing rapidly—reaching a total of $12 billion in the first five months of 2023 alone. Venture capital and other private external investments in generative AI increased by an average compound growth rate of 74 percent annually from 2017 to 2022. During the same period, investments in artificial intelligence overall rose annually by 29 percent, albeit from a higher base.

The rush to throw money at all things generative AI reflects how quickly its capabilities have developed. ChatGPT was released in November 2022. Four months later, OpenAI released a new large language model, or LLM, called GPT-4 with markedly improved capabilities.[2] Similarly, by May 2023, Anthropic's generative AI, Claude, was able to process 100,000 tokens of text, equal to about 75,000 words in a minute—the length of the average novel—compared with roughly 9,000 tokens when it was introduced in March 2023.[3] And in May 2023, Google announced several new features powered by generative AI, including Search Generative Experience and a new LLM called PaLM 2 that will power its Bard chatbot, among other Google products.[4]

From a geographic perspective, external private investment in generative AI, mostly from tech giants and venture capital firms, is largely concentrated in North America, reflecting the continent's current domination of the overall AI investment landscape. Generative AI–related companies based in the United States raised about $8 billion from 2020 to 2022, accounting for 75 percent of total investments in such companies during that period.[5]

Generative AI has stunned and excited the world with its potential for reshaping how knowledge work gets done in industries and business functions across the entire economy. Across functions such as sales and marketing, customer operations, and software development, it is poised to transform roles and boost performance. In the process, it could unlock trillions of dollars in value across sectors from banking to life sciences. We have used two overlapping lenses in this report to understand



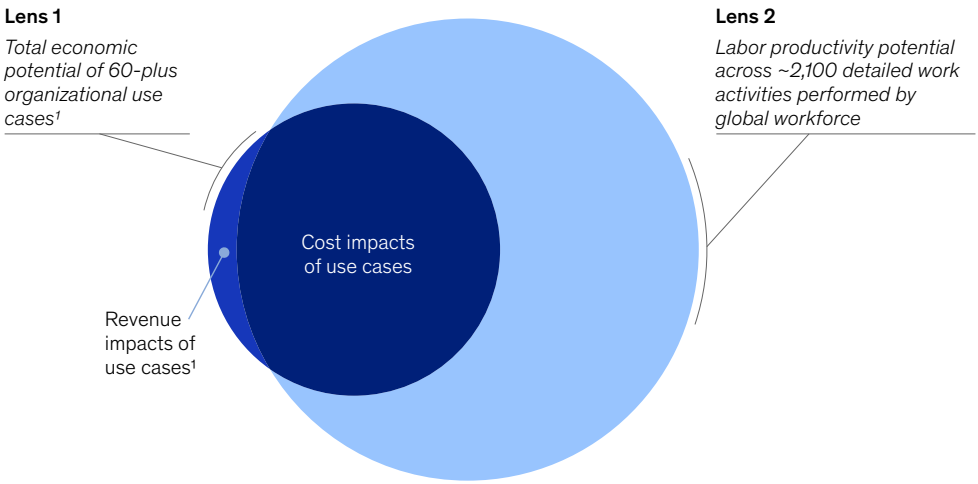
2

Generative AI use cases across functions and industries

the potential for generative AI to create value for companies and alter the workforce. The following sections share our initial findings.

Exhibit 1

The potential impact of generative AI can be evaluated through two lenses.



[1]For quantitative analysis, revenue impacts were recast as productivity increases on the corresponding spend in order to maintain comparability with cost impacts and not to assume additional growth in any particular market.

McKinsey & Company

Generative AI is a step change in the evolution of artificial intelligence. As companies rush to adapt and implement it, understanding the technology's potential to deliver value to the economy and society at large will help shape critical decisions. We have used two complementary lenses to determine where generative AI with its current capabilities could deliver the biggest value and how big that value could be (Exhibit 1).

The first lens scans use cases for generative AI that organizations could adopt. We define a "use case" as a targeted application of generative AI to a specific business challenge, resulting in one or more measurable outcomes. For example, a use case in marketing is the application of generative AI to generate creative content such as personalized emails, the measurable outcomes of which potentially include reductions in the cost of generating such content and increases in revenue from the enhanced effectiveness of higher-quality content at scale. We identified 63 generative AI use cases spanning 16 business functions that could deliver total value in the range of $2.6 trillion to $4.4 trillion in economic benefits annually when applied across industries.

That would add 15 to 40 percent to the $11.0 trillion to $17.7 trillion of economic value that we now estimate nongenerative artificial intelligence and analytics could unlock. (Our previous estimate from 2017 was that AI could deliver $9.5 trillion to $15.4 trillion in economic value.)

Our second lens complements the first by analyzing generative AI's potential impact on the work activities required in some 850 occupations. We modeled scenarios to estimate when generative AI could perform each of more than 2,100 "detailed work activities"—

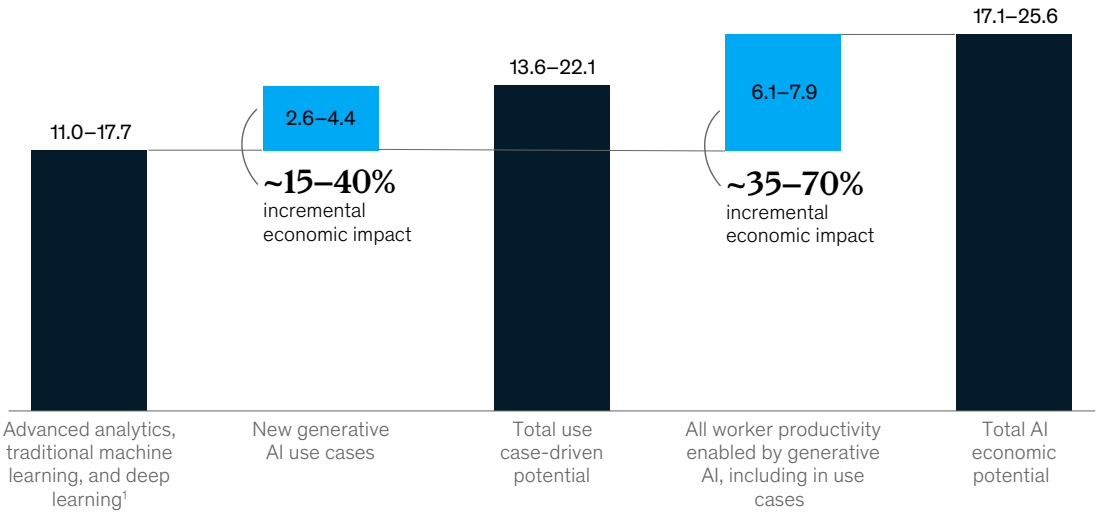
such as “communicating with others about operational plans or activities”—that make up those occupations across the world economy. This enables us to estimate how the current capabilities of generative AI could affect labor productivity across all work currently done by the global workforce.

Some of this impact will overlap with cost reductions in the use case analysis described above, which we assume are the result of improved labor productivity. Netting out this

Exhibit 2

Generative AI could create additional value potential above what could be unlocked by other AI and analytics.

AI’s potential impact on the global economy, $ trillion



[1]Updated use case estimates from "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.

McKinsey & Company

overlap, the total economic benefits of generative AI—including the major use cases we explored and the myriad increases in productivity that are likely to materialize when the technology is applied across knowledge workers' activities—amounts to $6.1 trillion to $7.9 trillion annually (Exhibit 2).

While generative AI is an exciting and rapidly advancing technology, the other applications of AI discussed in our previous report continue to account for the majority of the overall potential value of AI. Traditional advanced-analytics and machine learning algorithms are highly

Box 1

How we estimated the value potential of generative AI use cases

To assess the potential value of generative AI, we updated a proprietary McKinsey database of potential AI use cases and drew on the experience of more than 100 experts in industries and their business functions.[1] Our updates examined use cases of generative AI—specifically, how generative AI techniques (primarily transformer-based neural networks) can be used to solve problems not well addressed by previous technologies.

We analyzed only use cases for which generative AI could deliver a significant improvement in the outputs that drive key value. In particular, our estimates of the primary value the technology could unlock do not include use cases for which the sole benefit would be its ability to use natural language. For example, natural-language capabilities would be the key driver of value in a customer service use case but not in a use case optimizing a logistics network, where value primarily arises from quantitative analysis.

We then estimated the potential annual value of these generative AI use cases if they were adopted across the entire economy. For use cases aimed at increasing revenue, such as some of those in sales and marketing, we estimated the economy-wide value generative AI could deliver by increasing the productivity of sales and marketing expenditures.

Our estimates are based on the structure of the global economy in 2022 and do not consider the value generative AI could create if it produced entirely new product or service categories.

[1] "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.

effective at performing numerical and optimization tasks such as predictive modeling, and they continue to find new applications in a wide range of industries. However, as generative AI continues to develop and mature, it has the potential to open wholly new frontiers in creativity and innovation. It has already expanded the possibilities of what AI overall can achieve (see Box 1, "How we estimated the value potential of generative AI use cases").

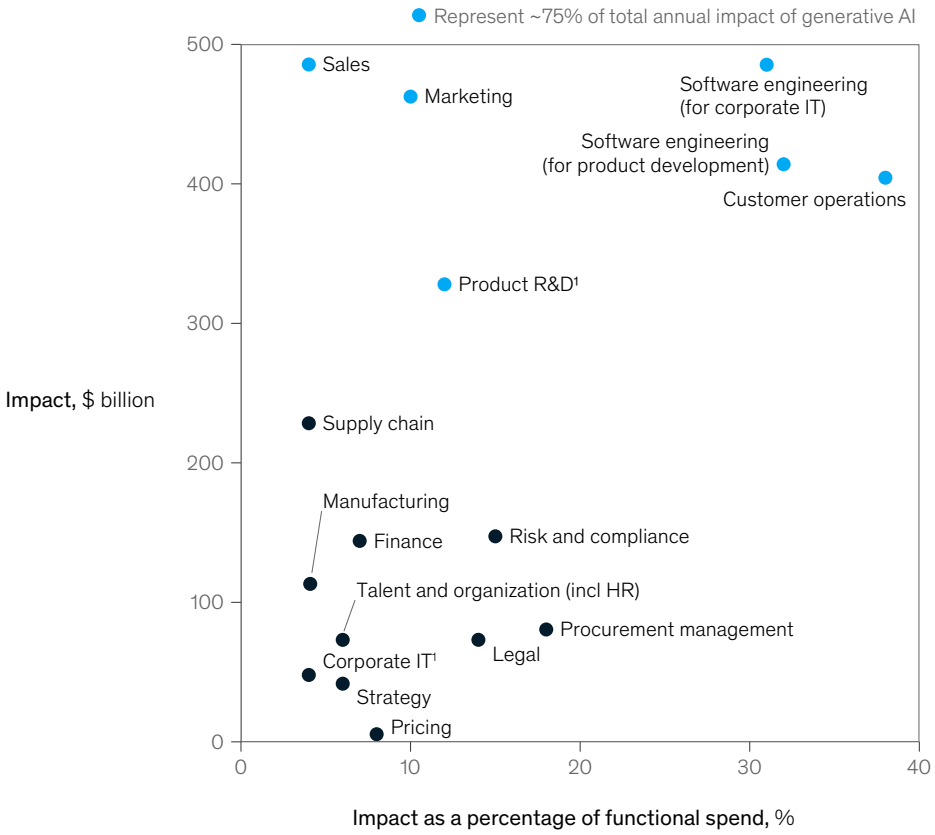
In this chapter, we highlight the value potential of generative AI across two dimensions: business function and modality.

Value potential by function

While generative AI could have an impact on most business functions, a few stand out when measured by the technology's impact as a share of functional cost (Exhibit 3). Our analysis of 16 business functions identified just four—customer operations, marketing and sales, software engineering, and research and development—that could account for approximately 75 percent of the total annual value from generative AI use cases.

Exhibit 3

Using generative AI in just a few functions could drive most of the technology's impact across potential corporate use cases.



Note: Impact is averaged.
[1]Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Notably, the potential value of using generative AI for several functions that were prominent in our previous sizing of AI use cases, including manufacturing and supply chain functions, is now much lower.[6] This is largely explained by the nature of generative AI use cases, which exclude most of the numerical and optimization applications that were the main value drivers for previous applications of AI.

Generative AI as a virtual expert

In addition to the potential value generative AI can deliver in function-specific use cases, the technology could drive value across an entire organization by revolutionizing internal knowledge management systems. Generative AI's impressive command of natural-language processing can help employees retrieve stored internal knowledge by formulating queries in the same way they might ask a human a question and engage in continuing dialogue. This could empower teams to quickly access relevant information, enabling them to rapidly make better-informed decisions and develop effective strategies.

In 2012, the McKinsey Global Institute (MGI) estimated that knowledge workers spent about a fifth of their time, or one day each workweek, searching for and gathering information. If generative AI could take on such tasks, increasing the efficiency and effectiveness of the workers doing them, the benefits would be huge. Such virtual expertise could rapidly "read" vast libraries of corporate information stored in natural language and quickly scan source material in dialogue with a human who helps fine-tune and tailor its research, a more scalable solution than hiring a team of human experts for the task.

Following are examples of how generative AI could produce operational benefits as a virtual expert in a handful of use cases.

In addition to the potential value generative AI can deliver in function-specific use cases, the technology could drive value across an entire organization by revolutionizing internal knowledge management systems.

Customer operations

Generative AI has the potential to revolutionize the entire customer operations function, improving the customer experience and agent productivity through digital self-service and enhancing and augmenting agent skills. The technology has already gained traction in customer service because of its ability to automate interactions with customers using natural language. Research found that at one company with 5,000 customer service agents, the application of generative AI increased issue resolution by 14 percent an hour and reduced the time spent handling an issue by 9 percent.[7] It also reduced agent attrition and requests to speak to a manager by 25 percent. Crucially, productivity and quality of service improved most among less-experienced agents, while the AI assistant did not increase—and sometimes decreased—the productivity and quality metrics of more highly skilled agents. This is because AI assistance helped less-experienced agents communicate using techniques similar to those of their higher-skilled counterparts.

The following are examples of the operational improvements generative AI can have for specific use cases:

— **Customer self-service.** Generative AI–fueled chatbots can give immediate and personalized responses to complex customer inquiries regardless of the language or location of the customer. By improving the quality and effectiveness of interactions via automated channels, generative AI could automate responses to a higher percentage of customer inquiries, enabling customer care teams to take on inquiries that can only be resolved by a human agent. Our research found that roughly half of customer contacts made by banking, telecommunications, and utilities companies in North America are already handled by machines, including but not exclusively AI. We estimate that generative AI could further reduce the volume of human-serviced contacts by up to 50 percent, depending on a company's existing level of automation.

— **Resolution during initial contact.** Generative AI can instantly retrieve data a company has on a specific customer, which can help a human customer service representative more successfully answer questions and resolve issues during an initial interaction.

— **Reduced response time.** Generative AI can cut the time a human sales representative spends responding to a customer by providing assistance in real time and recommending next steps.

— **Increased sales.** Because of its ability to rapidly process data on customers and their browsing histories, the technology can identify product suggestions and deals tailored to customer preferences. Additionally, generative AI can enhance quality assurance and coaching by gathering insights from customer conversations, determining what could be done better, and coaching agents.

We estimate that applying generative AI to customer care functions could increase productivity at a value ranging from 30 to 45 percent of current function costs.

Our analysis captures only the direct impact generative AI might have on the productivity of customer operations. It does not account for potential knock-on effects the technology may have on customer satisfaction and retention arising from an improved experience, including better understanding of the customer's context that can assist human agents in providing more personalized help and recommendations.

Marketing and sales

Generative AI has taken hold rapidly in marketing and sales functions, in which text-based communications and personalization at scale are driving forces. The technology can create personalized messages tailored to individual customer interests, preferences, and behaviors, as well as do tasks such as producing first drafts of brand advertising, headlines, slogans, social media posts, and product descriptions.

However, introducing generative AI to marketing functions requires careful consideration. For one thing, using mathematical models trained on publicly available data without sufficient safeguards against plagiarism, copyright violations, and branding recognition risks infringing on intellectual property rights. A virtual try-on application may produce biased representations of certain demographics because of limited or biased training data. Thus, significant human oversight is required for conceptual and strategic thinking specific to each company's needs.

Potential operational benefits from using generative AI for marketing include the following:

- **Efficient and effective content creation.** Generative AI could significantly reduce the time required for ideation and content drafting, saving valuable time and effort. It can also facilitate consistency across different pieces of content, ensuring a uniform brand voice, writing style, and format. Team members can collaborate via generative AI, which can integrate their ideas into a single cohesive piece. This would allow teams to significantly enhance personalization of marketing messages aimed at different customer segments, geographies, and demographics. Mass email campaigns can be instantly translated into as many languages as needed, with different imagery and messaging depending on the audience. Generative AI's ability to produce content with varying specifications could increase customer value, attraction, conversion, and retention over a lifetime and at a scale beyond what is currently possible through traditional techniques.
- **Enhanced use of data.** Generative AI could help marketing functions overcome the challenges of unstructured, inconsistent, and disconnected data—for example, from different databases—by interpreting abstract data sources such as text, image, and varying structures. It can help marketers better use data such as territory performance, synthesized customer feedback, and customer behavior to generate data-informed marketing strategies such as targeted customer profiles and channel recommendations. Such tools could identify and synthesize trends, key drivers, and market and product opportunities from unstructured data such as social media, news, academic research, and customer feedback.
- **SEO optimization.** Generative AI can help marketers achieve higher conversion and lower cost through search engine optimization (SEO) for marketing and sales technical components such as page titles, image tags, and URLs. It can synthesize key SEO tokens, support specialists in SEO digital content creation, and distribute targeted content to customers.
- **Product discovery and search personalization.** With generative AI, product discovery and search can be personalized with multimodal inputs from text, images and speech, and deep understanding of customer profiles. For example, technology can leverage individual user preferences, behavior, and purchase history to help customers discover the most

relevant products and generate personalized product descriptions. This would allow CPG, travel, and retail companies to improve their e-commerce sales by achieving higher website conversion rates.

We estimate that generative AI could increase the productivity of the marketing function with a value between 5 and 15 percent of total marketing spending.

Our analysis of the potential use of generative AI in marketing doesn't account for knock-on effects beyond the direct impacts on productivity. Generative AI–enabled synthesis could provide higher-quality data insights, leading to new ideas for marketing campaigns and better-targeted customer segments. Marketing functions could shift resources to producing higher-quality content for owned channels, potentially reducing spending on external channels and agencies.

Generative AI could also change the way both B2B and B2C companies approach sales. The following are two use cases for sales:

— **Increase probability of sale.** Generative AI could identify and prioritize sales leads by creating comprehensive consumer profiles from structured and unstructured data and suggesting actions to staff to improve client engagement at every point of contact. For example, generative AI could provide better information about client preferences, potentially improving close rates.

— **Improve lead development.** Generative AI could help sales representatives nurture leads by synthesizing relevant product sales information and customer profiles and creating discussion scripts to facilitate customer conversation, including up- and cross-selling talking points. It could also automate sales follow-ups and passively nurture leads until clients are ready for direct interaction with a human sales agent.

Our analysis suggests that implementing generative AI could increase sales productivity by approximately 3 to 5 percent of current global sales expenditures.

This analysis may not fully account for additional revenue that generative AI could bring to sales functions. For instance, generative AI's ability to identify leads and follow-up capabilities could uncover new leads and facilitate more effective outreach that would bring in additional revenue. Also, the time saved by sales representatives due to generative AI's capabilities could be invested in higher-quality customer interactions, resulting in increased sales success.

Generative AI as a virtual collaborator

In other cases, generative AI can drive value by working in partnership with workers, augmenting their work in ways that accelerate their productivity. Its ability to rapidly digest mountains of data and draw conclusions from it enables the technology to offer insights and options that can dramatically enhance knowledge work. This can significantly speed up the process of developing a product and allow employees to devote more time to higher-impact tasks.

Generative AI could increase sales productivity by 3 to 5 percent of current global sales expenditures.

Software engineering
Treating computer languages as just another language opens new possibilities for software engineering. Software engineers can use generative AI in pair programming and to do augmented coding and train LLMs to develop applications that generate code when given a natural-language prompt describing what that code should do.

Software engineering is a significant function in most companies, and it continues to grow as all large companies, not just tech titans, embed software in a wide array of products and services. For example, much of the value of new vehicles comes from digital features such as adaptive cruise control, parking assistance, and IoT connectivity.

According to our analysis, the direct impact of AI on the productivity of software engineering could range from 20 to 45 percent of current annual spending on the function. This value would arise primarily from reducing time spent on certain activities, such as generating initial code drafts, code correction and refactoring, root-cause analysis, and generating new system designs. By accelerating the coding process, generative AI could push the skill sets and capabilities needed in software engineering toward code and architecture design. One study found that software developers using Microsoft's GitHub Copilot completed tasks 56 percent faster than those not using the tool.[8] An internal McKinsey empirical study of software engineering teams found those who were trained to use generative AI tools rapidly reduced the time needed to generate and refactor code—and engineers also reported a better work experience, citing improvements in happiness, flow, and fulfillment.

Our analysis did not account for the increase in application quality and the resulting boost in productivity that generative AI could bring by improving code or enhancing IT architecture—which can improve productivity across the IT value chain. However, the quality of IT architecture still largely depends on software architects, rather than on initial drafts that generative AI's current capabilities allow it to produce.

Large technology companies are already selling generative AI for software engineering, including GitHub Copilot, which is now integrated with OpenAI's GPT-4, and Replit, used by more than 20 million coders.[9]

Product R&D

Generative AI's potential in R&D is perhaps less well recognized than its potential in other business functions. Still, our research indicates the technology could deliver productivity with a value ranging from 10 to 15 percent of overall R&D costs.

For example, the life sciences and chemical industries have begun using generative AI foundation models in their R&D for what is known as generative design. Foundation models can generate candidate molecules, accelerating the process of developing new drugs and materials. Entos, a biotech pharmaceutical company, has paired generative AI with automated synthetic development tools to design small-molecule therapeutics. But the same principles can be applied to the design of many other products, including larger-scale physical products and electrical circuits, among others.

While other generative design techniques have already unlocked some of the potential to apply AI in R&D, their cost and data requirements, such as the use of "traditional" machine learning, can limit their application. Pretrained foundation models that underpin generative AI, or models that have been enhanced with fine-tuning, have much broader areas of application than models optimized for a single task. They can therefore accelerate time to market and broaden the types of products to which generative design can be applied. For now, however, foundation models lack the capabilities to help design products across all industries.

In addition to the productivity gains that result from being able to quickly produce candidate designs, generative design can also enable improvements in the designs themselves, as in the following examples of the operational improvements generative AI could bring:

— **Enhanced design.** Generative AI can help product designers reduce costs by selecting and using materials more efficiently. It can also optimize designs for manufacturing, which can lead to cost reductions in logistics and production.

— **Improved product testing and quality.** Using generative AI in generative design can produce a higher-quality product, resulting in increased attractiveness and market appeal. Generative AI can help to reduce testing time of complex systems and accelerate trial phases involving customer testing through its ability to draft scenarios and profile testing candidates.

We also identified a new R&D use case for nongenerative AI: deep learning surrogates, the use of which has grown since our earlier research, can be paired with generative AI to produce even greater benefits (see Box 2, "Deep learning surrogates"). To be sure, integration will require the development of specific solutions, but the value could be significant because deep learning surrogates have the potential to accelerate the testing of designs proposed by generative AI.

While we have estimated the potential direct impacts of generative AI on the R&D function, we did not attempt to estimate the technology's potential to create entirely novel product categories. These are the types of innovations that can produce step changes not only in the performance of individual companies but in economic growth overall.

Value potential by modality

Technology has revolutionized the way we conduct business, and text-based AI is on the frontier of this change. Indeed, text-based data is plentiful, accessible, and easily processed and analyzed at large scale by LLMs, which has prompted a strong emphasis on them in the initial stages of generative AI development. The current investment landscape in generative AI is also heavily focused on text-based applications such as chatbots, virtual assistants, and language translation. However, we estimate that almost one-fifth of the value that generative AI can unlock across our use cases would take advantage of multimodal capabilities beyond text to text.

Box 2

Deep learning surrogates

Product design in industries producing physical products often involves physics-based virtual simulations such as computational fluid dynamics (CFD) and finite element analysis (FEA). Although they are faster than actual physical testing, these techniques can be time- and resource-intensive, especially for designing complex parts—running CFD simulations on graphics processing units can take hours. And these techniques are even more complex and compute-intensive when they involve simulations coupled across multiple disciplines (for example, physical stress and temperature distribution), which is sometimes called multiphysics.

Deep learning applications are now revolutionizing the virtual testing phase of the R&D process by using deep learning models to emulate (multi)physics-based simulations at higher speeds and lower costs. Instead of taking hours to run physics-based models, these deep learning surrogates can produce the results of simulations in just a few seconds, allowing researchers to test many more designs and enabling faster decision making on products and designs.

While most of generative AI's initial traction has been in text-based use cases, recent advances in generative AI have also led to breakthroughs in image generation, as OpenAI's DALL·E and Stable Diffusion have so amply illustrated, and much progress is being made in audio, including voice and music, and video. These capabilities have obvious applications in marketing for generating advertising materials and other marketing content, and these technologies are already being applied in media industries, including game design. Indeed, some of these examples challenge existing business models around talent, monetization, and intellectual property.[10]

The multimodal capabilities of generative AI could also be used effectively in R&D. Generative AI systems could create first drafts of circuit designs, architectural drawings, structural engineering designs, and thermal designs based on prompts that describe requirements for a product. Achieving this will require training foundation models in these domains (think of LLMs trained on "design languages"). Once trained, such foundation models could increase productivity on a similar magnitude to software development.

Value potential by industry

Across the 63 use cases we analyzed, generative AI has the potential to generate $2.6 trillion to $4.4 trillion in value across industries. Its precise impact will depend on a variety of factors, such as the mix and importance of different functions, as well as the scale of an industry's revenue (Exhibit 4).

Exhibit 4

Generative AI use cases will have different impacts on business functions across industries.

Generative AI productivity impact by business functions[1]

Low impact — High impact

	Total, % of industry revenue	Total, $ billion
Administrative and professional services	0.9–1.4	150–250
Advanced electronics and semiconductors	1.3–2.3	100–170
Advanced manufacturing[3]	1.4–2.4	170–290
Agriculture	0.6–1.0	40–70
Banking	2.8–4.7	200–340
Basic materials	0.7–1.2	120–200
Chemical	0.8–1.3	80–140
Construction	0.7–1.2	90–150
Consumer packaged goods	1.4–2.3	160–270
Education	2.2–4.0	120–230
Energy	1.0–1.6	150–240
Healthcare	1.8–3.2	150–260
High tech	4.8–9.3	240–460
Insurance	1.8–2.8	50–70
Media and entertainment	1.8–3.1	80–130
Pharmaceuticals and medical products	2.6–4.5	60–110
Public and social sector	0.5–0.9	70–110
Real estate	1.0–1.7	110–180
Retail[4]	1.2–1.9	240–390
Telecommunications	2.3–3.7	60–100
Travel, transport, and logistics	1.2–2.0	180–300
		2,600–4,400

Note: Figures may not sum to 100%, because of rounding.
[1]Excludes implementation costs (eg, training, licenses).
[2]Excluding software engineering.
[3]Includes aerospace, defense, and auto manufacturing.
[4]Including auto retail.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

For example, our analysis estimates generative AI could contribute roughly $310 billion in additional value for the retail industry (including auto dealerships) by boosting performance in functions such as marketing and customer interactions. By comparison, the bulk of potential value in high tech comes from generative AI's ability to increase the speed and efficiency of software development.

3

The generative AI future of work: Impacts on work activities, economic growth, and productivity

Technology has been changing the anatomy of work for decades. Over the years, machines have given human workers various “superpowers”; for instance, industrial-age machines enabled workers to accomplish physical tasks beyond the capabilities of their own bodies. More recently, computers have enabled knowledge workers to perform calculations that would have taken years to do manually.

These examples illustrate how technology can augment work through the automation of individual activities that workers would have otherwise had to do themselves. At a conceptual level, the application of generative AI may follow the same pattern in the modern workplace, although as we show later in this chapter, the types of activities that generative AI could affect, and the types of occupations with activities that could change, will likely be different as a result of this technology than for older technologies.

The McKinsey Global Institute began analyzing the impact of technological automation of work activities and modeling scenarios of adoption in 2017. At that time, we estimated that workers spent half of their time on activities that had the potential to be automated by adapting technology that existed at that time, or what we call technical automation potential. We also modeled a range of potential scenarios for the pace at which these technologies could be adopted and affect work activities throughout the global economy.

Technology adoption at scale does not occur overnight. The potential of technological capabilities in a lab does not necessarily mean they can be immediately integrated into a solution that automates a specific work activity—developing such solutions takes time. Even when such a solution is developed, it might not be economically feasible to use if its costs exceed those of human labor. Additionally, even if economic incentives for deployment exist, it takes time for adoption to spread across the global economy. Hence, our adoption scenarios, which consider these factors together with the technical automation potential, provide a sense of the pace and scale at which workers' activities could shift over time.

Large-scale shifts in the mix of work activities and occupations are not unprecedented. Consider the work of a farmer today compared with what a farmer did just a few short years ago. Many farmers now access market information on mobile phones to determine when and where to sell their crops or download sophisticated modeling of weather patterns. From a more macro perspective, agricultural employment in China went from an 82 percent share of all workers in 1962 to 13 percent in 2013. Labor markets are also dynamic: millions of people leave their jobs every month in the United States.[11] But this does not minimize the challenges faced by individual workers whose lives are upended by these shifts, or the organizational or societal challenges of ensuring that workers have the skills to take on the work that will be in demand and that their incomes are sufficient to grow their standards of living.

Also, demographics have made such shifts in activities a necessity from a macroeconomic perspective. An economic growth gap has opened as a result of the slowing growth of the world's workforce. In some major countries, workforces have shrunk because populations are aging. Labor productivity will have to accelerate to achieve economic growth and enhance prosperity.

The analyses in this paper incorporate the potential impact of generative AI on today's work activities. The new capabilities of generative AI, combined with previous technologies and integrated into corporate operations around the world, could accelerate the potential for technical automation of individual activities and the adoption of technologies that augment the capabilities of the workforce. They could also have an impact on knowledge workers whose activities were not expected to shift as a result of these technologies until later in the future.

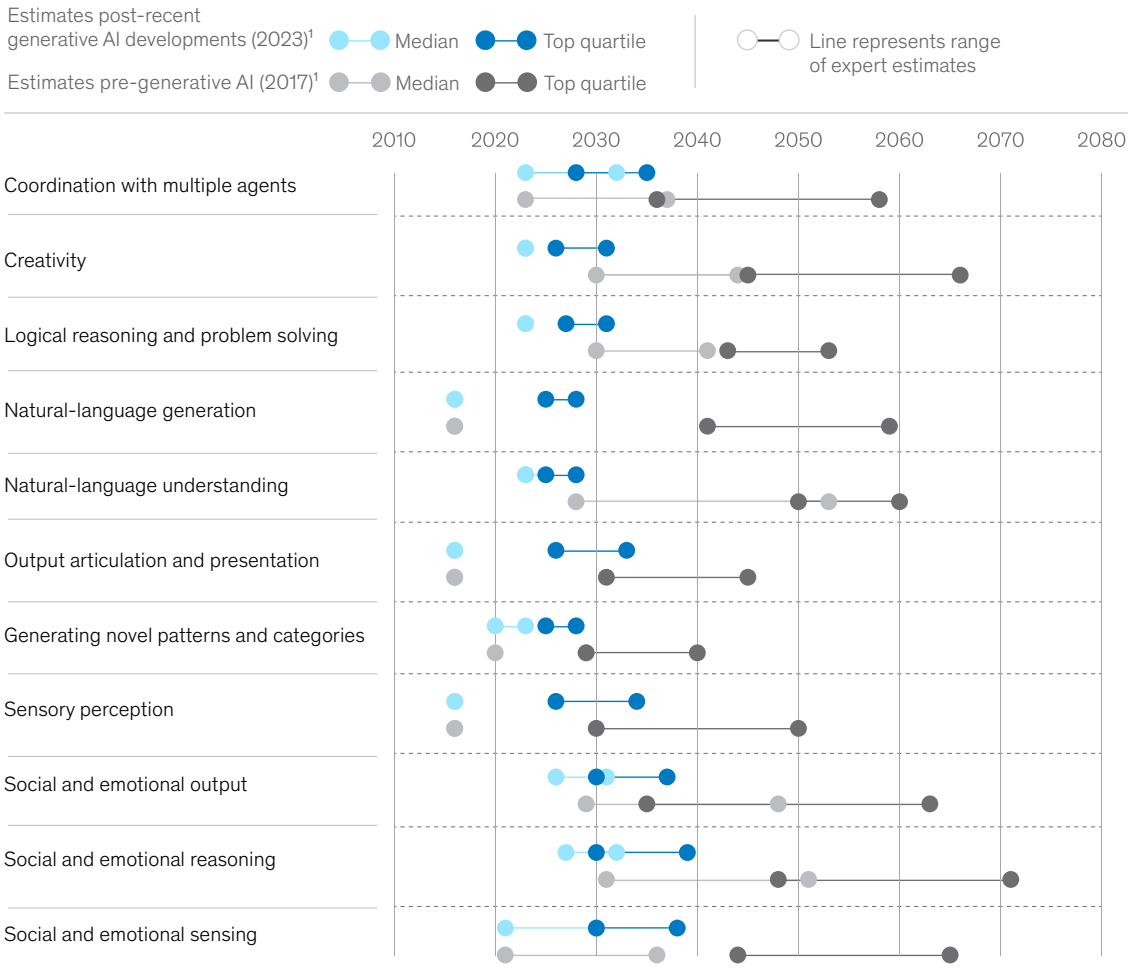
Accelerating the technical potential to transform knowledge work

Based on developments in generative AI, technology performance is now expected to match median human performance and reach top-quartile human performance earlier than previously estimated across a wide range of capabilities (Exhibit 5). For example, MGI previously identified 2027 as the earliest year when median human performance for natural-language understanding might be achieved in technology, but in this new analysis, the corresponding point is 2023.

Exhibit 5

As a result of generative AI, experts assess that technology could achieve human-level performance in some technical capabilities sooner than previously thought.

Technical capabilities, level of human performance achievable by technology



[1]Comparison made on the business-related tasks required from human workers. Please refer to technical appendix for detailed view of performance rating methodology.
Source: McKinsey Global Institute occupation database; McKinsey analysis

McKinsey & Company

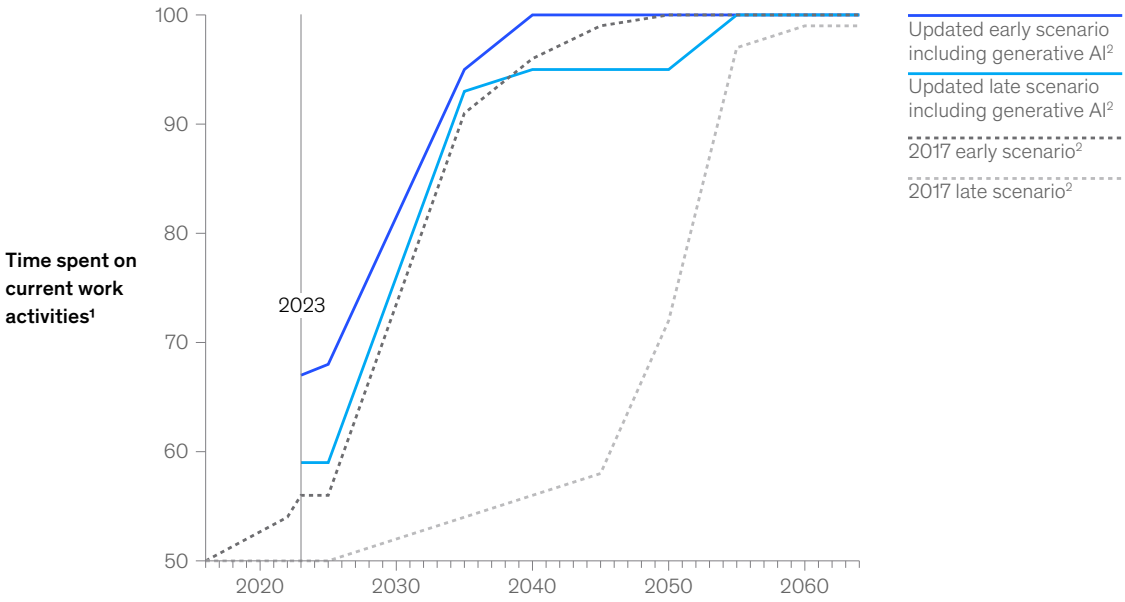
As a result of these reassessments of technology capabilities due to generative AI, the total percentage of hours that could theoretically be automated by integrating technologies that exist today has increased from about 50 percent to 60 to 70 percent. The technical potential curve is quite steep because of the acceleration in generative AI's natural-language capabilities (Exhibit 6).

Interestingly, the range of times between the early and late scenarios has compressed compared with the expert assessments in 2017, reflecting a greater confidence that higher levels of technological capabilities will arrive by certain time periods.

Exhibit 6

The advent of generative AI has pulled forward the potential for technical automation.

Technical automation potentials by scenario, %



[1]Includes data from 47 countries, representing about 80% of employment across the world. 2017 estimates are based on the activity and occupation mix from 2016. Scenarios including generative AI are based on the 2021 activity and occupation mix.
[2]Early and late scenarios reflect the ranges provided by experts (see Exhibit 6).
Source: McKinsey Global Institute analysis

McKinsey & Company

Generative AI could propel higher productivity growth

Global economic growth was slower from 2012 to 2022 than in the two preceding decades.[12] Although the COVID-19 pandemic was a significant factor, long-term structural challenges—including declining birth rates and aging populations—are ongoing obstacles to growth.

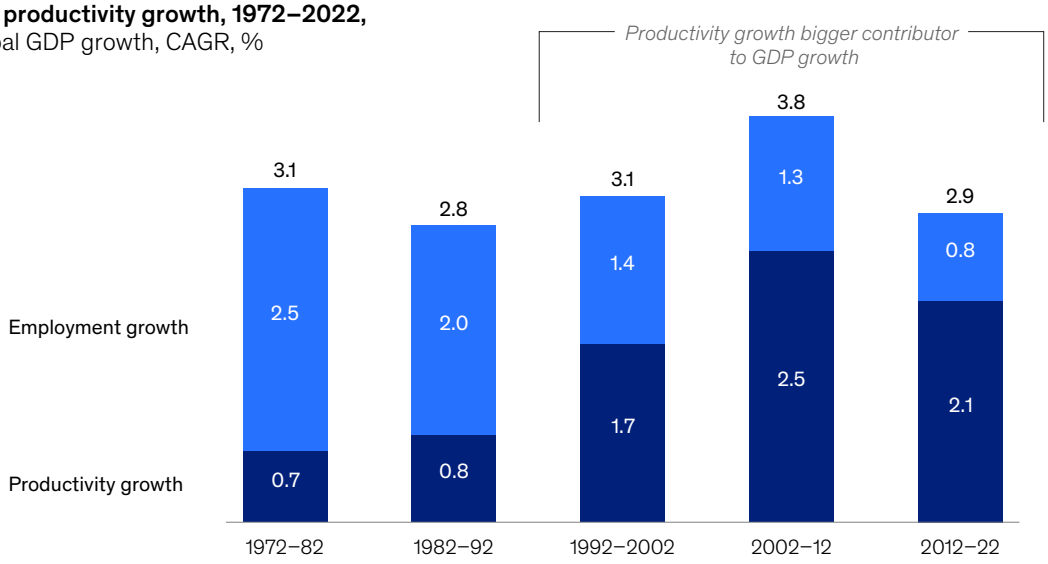
Declining employment is among those obstacles. Compound annual growth in the total number of workers worldwide slowed from 2.5 percent in 1972–82 to just 0.8 percent in 2012–22, largely because of aging. In many large countries, the size of the workforce is already declining.[13]

Productivity, which measures output relative to input, or the value of goods and services produced divided by the amount of labor, capital, and other resources required to produce them, was the main engine of economic growth in the three decades from 1992 to 2022 (Exhibit 7). However, since then, productivity growth has slowed in tandem with slowing employment growth, confounding economists and policy makers.[14]

Exhibit 7

Productivity growth, the main engine of GDP growth over the past 30 years, slowed down in the past decade.

Real GDP growth contribution of employment and productivity growth, 1972–2022,
global GDP growth, CAGR, %



Source: Conference Board Total Economy database; McKinsey Global Institute analysis

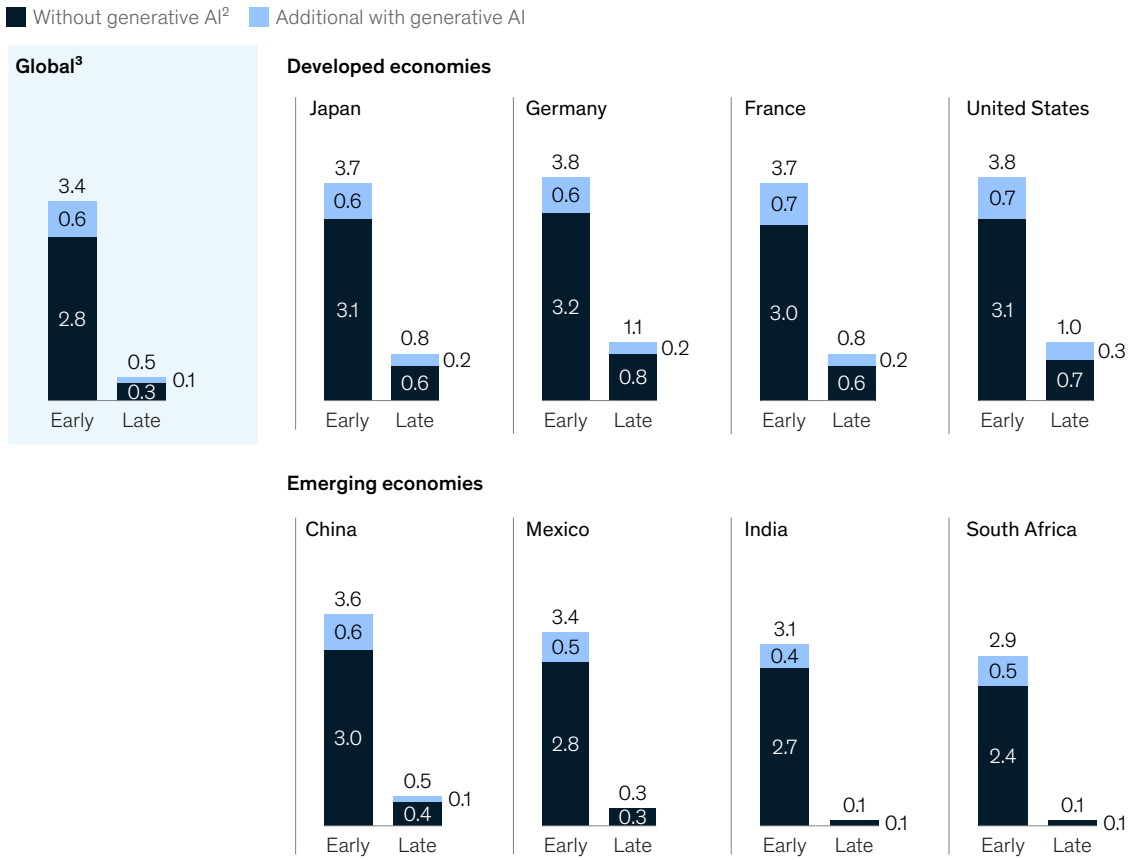
McKinsey & Company

The deployment of generative AI and other technologies could help accelerate productivity growth, partially compensating for declining employment growth and enabling overall economic growth. Based on our estimates, the automation of individual work activities enabled by these technologies could provide the global economy with an annual productivity boost of 0.5 to 3.4 percent from 2023 to 2040 depending on the rate of automation adoption—with generative AI contributing to 0.1 to 0.6 percentage points of that growth—but only if individuals affected by the technology were to shift to other work activities that at least match their 2022 productivity levels (Exhibit 8). In some cases, workers will stay in the same occupations, but their mix of activities will shift; in others, workers will need to shift occupations.

Exhibit 8

Generative AI could contribute to productivity growth if labor hours can be redeployed effectively.

Productivity impact from automation by scenario, 2022–40, CAGR,[1] %



Note: Figures may not sum, because of rounding.
[1]Based on the assumption that automated work hours are reintegrated in work at productivity level of today.
[2]Previous assessment of work automation before the rise of generative AI.
[3]Based on 47 countries, representing about 80% of world employment.
Source: Conference Board Total Economy Database; Oxford Economics; McKinsey Global Institute analysis

McKinsey & Company

The capabilities of generative AI vastly expand the pool of work activities with the potential for technical automation. That in turn has sped up the pace at which automation may be deployed and expanded the types of workers who will experience its impact. Like other technologies, its ability to take on routine tasks and work can increase human productivity, which has grown at a below-average rate for almost 20 years.[15] It can also offset the impact of aging, which is beginning to put a dent in workforce growth for many of the world's major economies. But to achieve these benefits, a significant number of workers will need to substantially change the work they do, either in their existing occupations or in new ones. They will also need support in making transitions to new activities.

4

Considerations for businesses and society

History has shown that new technologies have the potential to reshape societies. Artificial intelligence has already changed the way we live and work—for example, it can help our phones (mostly) understand what we say, or draft emails. Mostly, however, AI has remained behind the scenes, optimizing business processes or making recommendations about the next product to buy. The rapid development of generative AI is likely to significantly augment the impact of AI overall, generating trillions of dollars of additional value each year and transforming the nature of work.

But the technology could also deliver new and significant challenges. Stakeholders must act—and quickly, given the pace at which generative AI could be adopted—to prepare to address both the opportunities and the risks. Risks have already surfaced, including concerns about the content that generative AI systems produce: Will they infringe upon intellectual property due to "plagiarism" in the training data used to create foundation models? Will the answers that LLMs produce when questioned be accurate, and can they be explained? Will the content that generative AI creates be fair or biased in ways that users do not want by, say, producing content that reflects harmful stereotypes?

There are economic challenges too: the scale and the scope of the workforce transitions described in this report are considerable. In the midpoint adoption scenario, about a quarter to a third of work activities could change in the coming decade. The task before us is to manage the potential

positives and negatives of the technology simultaneously (for more about the potential risks of generative AI, see Box 3, "Using generative AI responsibly"). Here are some of the critical questions we will need to address while balancing our enthusiasm for the potential benefits of the technology with the new challenges it can introduce.

Companies and business leaders

How can companies move quickly to capture the potential value at stake highlighted in this report, while managing the risks that generative AI presents?

Box 3

Using generative AI responsibly

Generative AI poses a variety of risks. Stakeholders will want to address these risks from the start.

Fairness: Models may generate algorithmic bias due to imperfect training data or decisions made by the engineers developing the models.

Intellectual property (IP): Training data and model outputs can generate significant IP risks, including infringing on copyrighted, trademarked, patented, or otherwise legally protected materials. Even when using a provider's generative AI tool, organizations will need to understand what data went into training and how it's used in tool outputs.

Privacy: Privacy concerns could arise if users input information that later ends up in model outputs in a form that makes individuals identifiable. Generative AI could also be used to create and disseminate malicious content such as disinformation, deepfakes, and hate speech.

Security: Generative AI may be used by bad actors to accelerate the sophistication and speed of cyberattacks. It also can be manipulated to provide malicious outputs. For example, through a technique called prompt injection, a third party gives a model new instructions that trick the model into delivering an output unintended by the model producer and end user.

Explainability: Generative AI relies on neural networks with billions of parameters, challenging our ability to explain how any given answer is produced.

Reliability: Models can produce different answers to the same prompts, impeding the user's ability to assess the accuracy and reliability of outputs.

Organizational impact: Generative AI may significantly affect the workforce, and the impact on specific groups and local communities could be disproportionately negative.

Social and environmental impact: The development and training of foundation models may lead to detrimental social and environmental consequences, including an increase in carbon emissions (for example, training one large language model can emit about 315 tons of carbon dioxide).[1]

[1] Ananya Ganesh, Andrew McCallum, and Emma Strubell, "Energy and policy considerations for deep learning in NLP," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, June 5, 2019.

How will the mix of occupations and skills needed across a company's workforce be transformed by generative AI and other artificial intelligence over the coming years? How will a company enable these transitions in its hiring plans, retraining programs, and other aspects of human resources?

Do companies have a role to play in ensuring the technology is not deployed in "negative use cases" that could harm society?

How can businesses transparently share their experiences with scaling the use of generative AI within and across industries—and also with governments and society?

Policy makers
What will the future of work look like at the level of an economy in terms of occupations and skills? What does this mean for workforce planning?

How can workers be supported as their activities shift over time? What retraining programs can be put in place? What incentives are needed to support private companies as they invest in human capital? Are there earn-while-you-learn programs such as apprenticeships that could enable people to retrain while continuing to support themselves and their families?

What steps can policy makers take to prevent generative AI from being used in ways that harm society or vulnerable populations?

Can new policies be developed and existing policies amended to ensure human-centric AI development and deployment that includes human oversight and diverse perspectives and accounts for societal values?

Individuals as workers, consumers, and citizens
How concerned should individuals be about the advent of generative AI? While companies can assess how the technology will affect their bottom lines, where can citizens turn for accurate, unbiased information about how it will affect their lives and livelihoods?

How can individuals as workers and consumers balance the conveniences generative AI delivers with its impact in their workplaces?

Can citizens have a voice in the decisions that will shape the deployment and integration of generative AI into the fabric of their lives?

Technological innovation can inspire equal parts awe and concern. When that innovation seems to materialize fully formed and becomes widespread seemingly overnight, both responses can be amplified. The arrival of generative AI in the fall of 2022 was the most recent example of this phenomenon, due to its unexpectedly rapid adoption as well as the ensuing scramble among companies and consumers to deploy, integrate, and play with it.

All of us are at the beginning of a journey to understand this technology's power, reach, and capabilities. If the past eight months are any guide, the next several years will take us on a roller-coaster ride featuring fast-paced innovation and technological breakthroughs that force us to recalibrate our understanding of AI's impact on our work and our lives. It is important to properly understand this phenomenon and anticipate its impact. Given the speed of generative AI's deployment so far, the need to accelerate digital transformation and reskill labor forces is great.

These tools have the potential to create enormous value for the global economy at a time when it is pondering the huge costs of adapting to and mitigating climate change. At the same time, they also have the potential to be more destabilizing than previous generations of artificial intelligence. They are capable of that most human of abilities, language, which is a fundamental requirement of most work activities linked to expertise and knowledge as well as a skill that can be used to hurt feelings, create misunderstandings, obscure truth, and incite violence and even wars.

We hope this research has contributed to a better understanding of generative AI's capacity to add value to company operations and fuel economic growth and prosperity as well as its potential to dramatically transform how we work and our purpose in society. Companies, policy makers, consumers, and citizens can work together to ensure that generative AI delivers on its promise to create significant value while limiting its potential to upset lives and livelihoods. The time to act is now.[16]

Endnotes

1 Ryan Morrison, "Compute power is becoming a bottleneck for developing AI. Here's how you clear it.," *Tech Monitor*, updated March 17, 2023.

2 "Introducing ChatGPT," OpenAI, November 30, 2022; "GPT-4 is OpenAI's most advanced system, producing safer and more useful responses," OpenAI, accessed June 1, 2023.

3 "Introducing Claude," Anthropic PBC, March 14, 2023; "Introducing 100K Context Windows," Anthropic PBC, May 11, 2023.

4 Emma Roth, "The nine biggest announcements from Google I/O 2023," *The Verge*, May 10, 2023.

5 Pitchbook.

6 Ibid.

7 Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, *Generative AI at work*, National Bureau of Economic Research working paper number 31161, April 2023.

8 Peter Cihon et al., *The impact of AI on developer productivity: Evidence from GitHub Copilot*, Cornell University arXiv software engineering working paper, arXiv:2302.06590, February 13, 2023.

9 Michael Nuñez, "Google and Replit join forces to challenge Microsoft in coding tools," VentureBeat, March 28, 2023.

10 Joe Coscarelli, "An A.I. hit of fake 'Drake' and 'The Weeknd' rattles the music world," New York Times, updated April 24, 2023.

11 "Job openings and labor turnover survey," US Bureau of Labor Statistics, accessed June 6, 2023.

12 *Global economic prospects*, World Bank, January 2023.

13 Yaron Shamir, "Three factors contributing to fewer people in the workforce," *Forbes*, April 7, 2022.

14 "The U.S. productivity slowdown: an economy-wide and industry-level analysis," Monthly Labor Review, US Bureau of Labor Statistics, April 2021; Kweilin Ellingrud, "Turning around the productivity slowdown," McKinsey Global Institute, September 13, 2022.

15 "Rekindling US productivity for a new era," McKinsey Global Institute, February 16, 2023.

16 The research, analysis, and writing in this report was entirely done by humans.

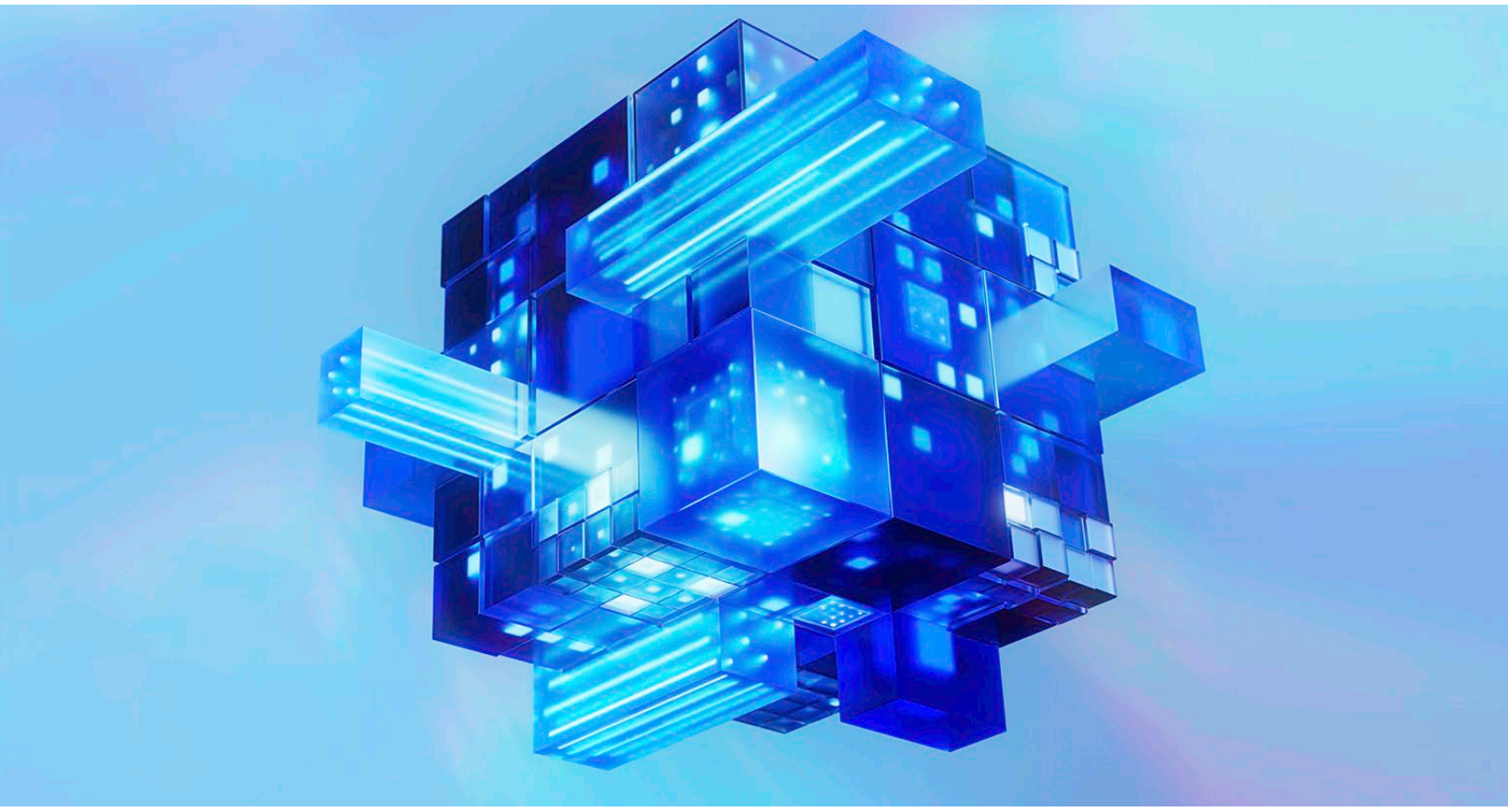
The research underpinning this report was led by **Michael Chui,** an MGI partner in McKinsey's Bay Area office; **Eric Hazan,** a senior partner in the Paris office; **Roger Roberts,** a partner in the Bay Area office; **Alex Singla,** a senior partner in the Chicago office; **Kate Smaje** and **Alexander Sukharevsky,** senior partners in the London office; **Lareina Yee,** a senior partner in the Bay Area office; and **Rodney Zemmel,** a senior partner in the New York office.

QuantumBlack
AI by McKinsey

Making the most of the generative AI opportunity: Six questions for CEOs

As corporate leaders navigate the new gen AI era, they can begin to lay out their road map and strategy by pondering a series of fundamental questions.

This article is a collaborative effort by Ben Ellencweig, Dana Maor, Alex Singla, Alexander Sukharevsky, Lareina Yee, and Rodney Zemmel, representing views from QuantumBlack, AI by McKinsey.



February 2024

Generative AI (gen AI) has taken the world by storm, altering our understanding of the possible. Creating next-era fashion collections in a few clicks, engaging customers with hyper-personalized offerings, and collapsing years of tedious drug discovery work into a few months—suddenly, all that and more seems within reach. As in the early days of breakthroughs like blockchain and the Internet itself, gen AI has sparked a debate between those who believe the technology will reshape the way we work and live and those who see gen AI as the next NFT moment, soaring briefly and failing to deliver on its promise, as nonfungible tokens did earlier in this decade.

So how much of today's excitement about gen AI reflects reality, and how much is myth? McKinsey estimates that the technology will open a new era of productivity and growth that could create $2.6 billion to $4.4 trillion of additional value.[1] In the telecom space alone, the impact of new gen AI use cases is expected to be in the range of $60 billion to $100 billion.

For CEOs seeking to unlock this upside, the key is to understand how this value will materialize and over what period, as well as where to invest their resources. There are no right answers, at least not yet. We are still in the technology's post-awareness, pre-deployment phase, with most software engineers having only recently gained access to gen AI tools. But based on our experience working with clients over the past 15 months, we find that CEOs can better formulate a strategy if they consider six essential questions about gen AI:

1. Is the opportunity significantly larger than AI?
2. Are we ambitious enough with gen AI?
3. Where is the money in the value chain?
4. Do we have the right talent in place?
5. What does it take to cross the "Death Valley" of scaling AI?
6. Are we thinking about risk in the right way?

Is the opportunity significantly larger than AI?

Over the past year, many of our client conversations and technology deployments have focused on gen AI. Despite its novelty, however, gen AI does not exist in a silo. Instead, it is simply the newest, if most powerful, iteration in the unfolding story of how artificial intelligence can boost productivity and innovation. We estimate that gen AI accounts for only 20 to 40 percent of AI's total value creation potential, with the remainder coming from traditional, or "analytical," AI applications, which have heretofore been less than fully deployed.

What's more, other important technology trends, such as Web 3.0 and augmented reality and virtual reality (AR/VR), are continuing to make progress in the shadow of gen AI. They will eventually get a strong footing over the next decade, with clear value creation potential for organizations. Hence, executives rethinking industries and business models should view the opportunity more broadly than gen AI or even all AI. A more effective approach is to consider how their organizations can capitalize on the confluence of emerging technology trends—a truly watershed moment akin to the simultaneous emergence of the first cloud, social network, and smartphone applications in 2017.

Are we ambitious enough with gen AI?

Gen AI has fascinated the world with jaw-dropping applications like ChatGPT and Pi, highlighting AI's transformative potential. Never before has technology pushed the art

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

of the possible so far ahead and so fast for non-technologists. As companies rush toward this technology, they are likelier to succeed if they solve for value creation versus simply checking the box. This is particularly concerning in the telco space, where players have often expressed interest in exploring incremental productivity applications and less frequently turn their attention to reimagining their businesses through the lens of AI.

Beyond simply avoiding a rehashed discussion of tech versus telco and scaling use cases, senior executives can benefit from asking a weightier question: How do we reinvent our industry and business model by leveraging the disintermediation, radical cost curve shifts, and organic consumer acquisition opportunities that gen AI can provide? Moreover, the age of new platforms opens new opportunities, including the creation from scratch of hyperscalers or unicorn super apps. One only needs to consider the opportunity associated with natural-language virtual assistants and the disruption this could have on the current business context, from consumption to business model. Gen AI will reward the bold. Already, some 80 percent of today's most popular gen AI products come from new entrants,[2] with incumbents forced to play catch-up or otherwise find their edge to lead.

Where is the money in the value chain?

Gen AI is creating a frenzy among founders and investors, with a seemingly endless number of players entering the field. A closer look at leading gen AI players reveals a couple of winning plays that CEOs might use to separate their organization from the pack[3]:

— **Differentiate à la the fine-dining chef.** The ingredients of gen AI applications are not in and of themselves a source of competitive differentiation. Anyone can license the most powerful closed-source models, which contribute only about 15 percent of the value of gen AI applications. This suggests that the real value will be realized by those able to combine the best available technology with proprietary data. Telco leaders should reexamine their data asset portfolios with an eye toward designing features like unique consumer and distribution journeys, such as an always-on customer care assistant fine-tuned to each user profile and embedded across user channels. Indeed, this is exactly what people seem to be asking for: among the top 50 gen AI applications, consumers are paying for 90 percent of them, revenue per user is three times higher than that of other apps, and customer acquisition is mostly organic.

— **Find underserved segments of the value chain.** Gen AI models and applications get most of the attention from investors and organizations, but other critical segments of the gen AI value chain remain surprisingly underrated. From commercializing access to graphical processing units (GPUs) to providing data cleaning, augmentation, or risk management solutions, the opportunities are plentiful. We see this play already taking place in the data center space, with investors exploring acquisitions to supply an accelerating demand for workloads. Fortunately, many opportunities still exist for organizations to gain first-mover advantages in the gen AI market. In fact, in most product categories, the gap between the top two players is only two times, making it easier for new entrants to establish themselves as leaders in the field.

Do we have the right talent in place?

Our research shows that gen AI is expected to supercharge automation, affecting up to 60 percent of work activities over the next 20 years. This impact should not be surprising; a gen AI model can analyze in an hour more data than a human can in ten lifetimes. But will AI replace us all or turn us into automatons?

[2] Olivia Moore, "How are consumers using generative AI?" Andreessen Horowitz, Sept. 13, 2023; By some estimates, gen AI start-ups alone have already generated more than $1 billion of software-as-a-service revenue.
[3] Ninety percent of these companies are already monetizing their offerings with more than three times the average revenue per user than incumbents.

Those concerns seem overwrought, at least for now. The fact is, gen AI can deliver only if it is combined with exceptional human capital. Despite the power of gen AI, middling employees will produce middling results. Organizations must recruit, retain, and develop truly outstanding talent in both the technical and nontechnical spheres. With the right people in place, organizations truly could be on the verge of a new age of innovation.

What does it take to cross the 'Death Valley' of scaling AI?

Only one in ten AI use cases have been deployed in production,[4] so gen AI has arrived at a time when many leaders are disillusioned with the yet-unfulfilled promises of artificial intelligence. But AI does not have a technology problem; it has a design problem. To be effective, AI models require top-down focus, the right tech and people capabilities, proper data access, modular architecture, and effective change management. Only then can disparate AI-driven solutions work together continuously to create great customer and employee experiences, lower unit costs, and allow the organization to move faster than ever. Without external intervention or guidance, only about 3 percent of gen AI proofs of concept eventually scale.

Creating a digitally capable organization involves rewiring the way companies operate. This effort should be broad, covering six dimensions:

1. *business-led digital road map* that aligns the senior leadership team on the transformation vision, value, and strategy, which is focused on business reinvention

2. *talent with the right skills* and capabilities to execute and innovate in both the technical and business sides of the organization, including upskilling

3. *operating model* that increases the organization's metabolic rate by bringing together business and technology

4. *technology* that allows the organization to innovate faster and more easily—in particular, an IT architecture with a flexible orchestration layer

5. *data* that is continuously enriched and easy to consume across the organization to improve customer experiences and business performance

6. *adoption and scaling* of digital and AI solutions to optimize value capture by building new skills and leadership characteristics and by tightly managing the transformation progress and risks

Are we thinking about risk in the right way?

Discussions of gen AI risks are plentiful, but experience shows that most of these conversations need calibrating to ensure that organizations approach risk holistically and pragmatically. Current conversations about risk tend to focus on either short-term considerations (for example, customer experience and protection of intellectual property) or long-term, existential ones (whether artificial general intelligence will rule the world). Not enough focus is placed on intermediate risk, such as how companies can maintain the trust of their stakeholders in an AI-generated reality where seeing is no longer believing. Also, other categories of risks are simply not getting the attention they deserve. Little is said, for example, about organizations' environmental, social, and governance (ESG) risk, even though training a gen AI model consumes about a million liters of water for cooling.

[4] "The state of AI in 2022—and a half decade in review," McKinsey, December 6, 2022.

A more pragmatic perspective would be for CEOs to steer their organizations toward accepting risk as the reality of doing business with AI (for example, hallucination is just a feature of gen AI). Fortunately, risks can be managed. Plenty of banks, after all, deal with customer credit and other difficult types of risk daily and still manage to thrive. To navigate these uncharted waters, organizations should set up cross-functional teams to cover their specific risk concerns (for example, regulatory, ethical, cyber, IP, and societal risks), establish ethical principles and guidelines for gen AI use, and establish continuous monitoring for gen AI systems to address risk dynamically.

An honest and thorough examination of these six questions can lay the foundation of a comprehensive gen AI strategy—one that truly focuses on how the technology can transform an organization or an entire industry. These conversations will not necessarily be easy, which makes it essential that they be led by CEOs. Perhaps most of all, it is advantageous to think big. A road map, after all, can lead to different destinations. Where do you want your company to land?

Ben Ellencweig is a senior partner in McKinsey's Stamford office, **Dana Maor** is a senior partner in the Tel Aviv office, and **Lareina Yee,** a senior partner and chair of the McKinsey Technology Council, is based in the Bay Area–San Francisco office. **Alex Singla,** a senior partner in the Chicago office, and **Alexander Sukharevsky,** a senior partner in the London office, are managing partners of QuantumBlack, AI by McKinsey. **Rodney Zemmel,** a senior partner and managing partner of McKinsey Digital, is based in the New York office.

2

Sector view: Telecom operators

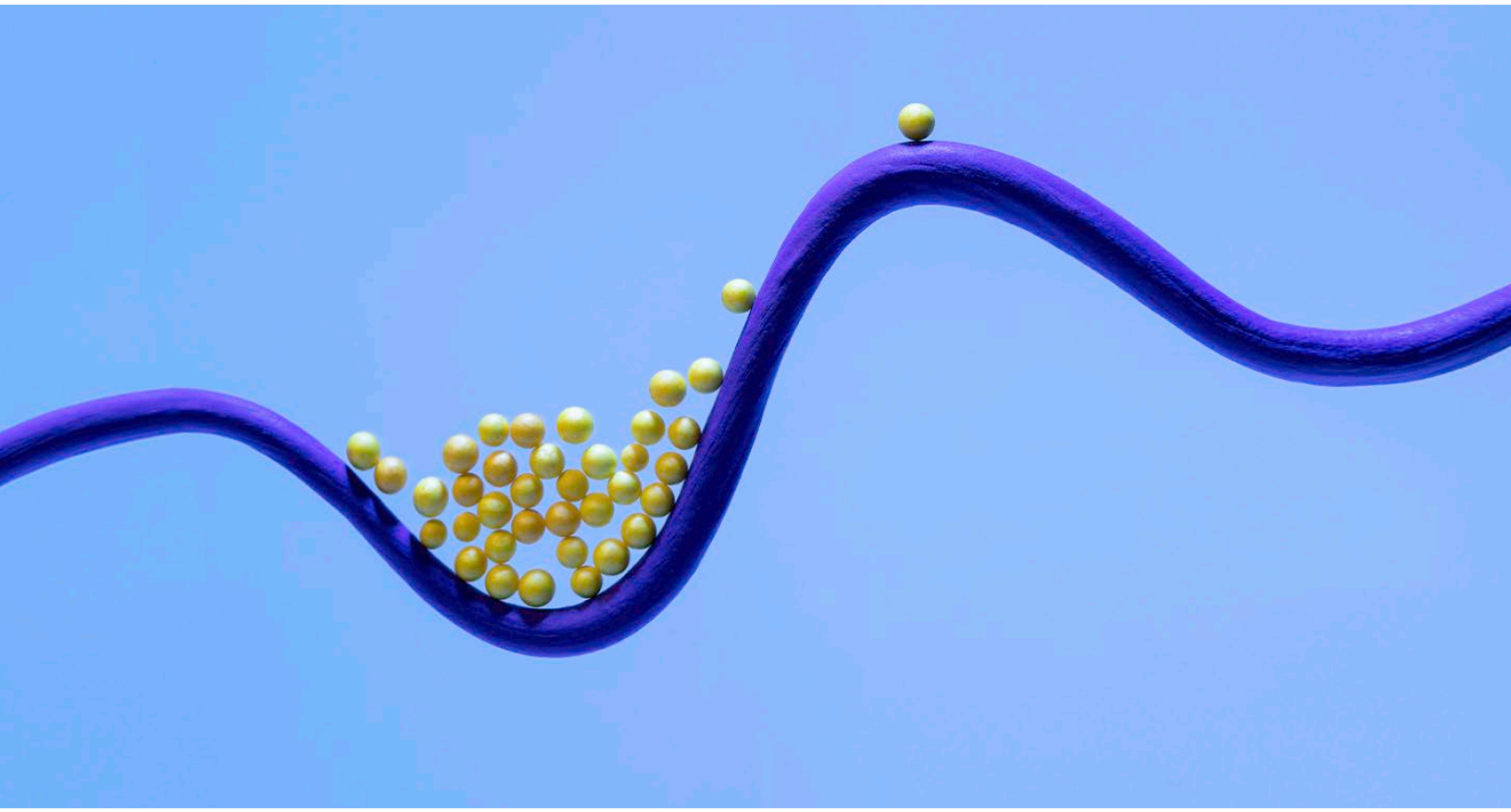
McKinsey & Company

Technology, Media & Telecommunications Practice

The AI-native telco: Radical transformation to thrive in turbulent times

Artificial intelligence, when deployed at scale, can help telcos protect core revenues and drive margin growth. But capturing this opportunity will require a wholly different approach.

This article is a collaborative effort by Joshan Abraham, Jorge Amar, Yuval Atsmon, Miguel Frade, and Tomás Lajous, representing views from McKinsey's Technology, Media & Telecommunications Practice.



February 2023

Artificial intelligence (AI) is unlocking use cases that are transforming industries across a wide swath of the world's economy. From infrastructure that "self-heals" to radically reimagined (and touchless) customer service and experience; from large scale hyperpersonalization to automatically created marketing messages and images leveraging Generative AI tools like ChatGPT—it is all a reality today. These AI solutions can powerfully augment and sometimes radically outperform most traditional business roles.

The impact from these solutions is becoming evident. AI leaders—the top quintile of companies that have taken the McKinsey Analytics Quotient assessment—have experienced a five-year revenue CAGR that is 2.1 times higher than that of peers and a total return to shareholders that is 2.5 times larger.

Given the numerous challenges the telecom industry has faced in recent years, such as flagging revenues and ROIC, one might expect the industry would have already adopted a full transition to this technology. Yet, based on our experience with operators across the world, telcos have yet to fully embrace AI and an AI-focused mindset. Instead, models are developed once and not enhanced as the business context evolves. Machine learning (ML) is in name only, limiting the ability of the system to improve from experience. Most regrettably, AI investments are often not aligned with top-level management priorities; lacking that sponsorship, AI deployments stall, investment in technical talent withers, and the technology remains immature.

Contrast this disjointed state of affairs with an AI-native organization. Here, AI is viewed as a core competency that powers decision making across all departments and organization layers. AI investments are required to enable most C-level priorities such as more personalized recommendations for customers and faster speed of answer in call centers. Top executives serve as champions of critical AI initiatives. Data and AI capabilities are managed as products, built for scalability and reusability. AI product managers, even those working on foundational products, are celebrated for the benefits they generate for the organization.

Reaching this state of AI maturity is no easy task, but it is certainly within the reach of telcos. Indeed, with all the pressures they face, embracing large-scale deployment of AI and transitioning to being AI-native organizations could be key to driving growth and renewal. Telcos that are starting to recognize this is nonnegotiable are scaling AI investments as the business impact generated by the technology materializes.

While isolated applications of the technology can help individual departments improve, it's AI connected holistically at all levels and departments that will be key to protecting core revenue and driving margin growth in even the most difficult of environments. Imagine the following not-so-distant scenarios:

— *Customer focused:* Sarah, a New Yorker, is a high average revenue per user (ARPU) customer. Aware that Sarah spends half of her phone usage time on fitness apps, the AI creates an enticing customized upgrade offer that includes a six-month credit applicable to her favorite fitness subscription and NYC-specific perks, such as a ticket to an upcoming concert sponsored by the operator. Knowing Sarah's high digital propensity[1], the AI makes the offer available to her as a digital-only promotion.

— *Employee focused:* When Trevor, an associate in a telco mall store, logs in at the start of his shift, he receives a celebratory notification congratulating him on his high-quality interactions with customers the previous day. And because the AI detected that Trevor is underperforming peers in accessory and device protection attach rates, he receives a notification pointing him to coaching resources specifically created to enhance performance in those metrics.

[1] Preference to transact and engage in digital channels, such as websites and mobile apps.

— *Infrastructure focused:* Lucile, director of a capital planning team, uses AI to inform highly targeted network investment decisions based on a granular understanding of customer-level network experience scores strongly correlated to commercial outcomes (for example, churn). The AI provides tactical recommendations of what and where to build based on where customers use the network and on automatically computed thresholds after which new investments have marginal impact on experience and commercial outcomes for the operator.

How these possibilities could become reality is critical to consider, especially given that most telcos currently deploy AI in limited ways that will not drive sustainable, at-scale success.

Why now? The case for becoming AI native

Factors supporting this move for telcos include the following:

— *Increasing accessibility of leading AI technology:* AI-native organizations like Meta continue to grow the open-source ecosystem by making new programming languages, datasets, and algorithms widely available. In parallel, cloud providers have developed multiple quick-to-deploy machine-learning APIs like Google Cloud's Natural Language API. Generative AI solutions, such as ChatGPT, that are capable of creating engaging responses to human queries are also accessible through API. These two factors, coupled with dropping costs of data processing and storage, make AI increasingly easier for organizations to leverage.

— *Rapid explosion of usable data:* Operators can collect, structure, and use significantly more data directly than ever before. This information includes data flows from individualized app usage patterns, site-specific customer experience scores, and what can be purchased or shared from partners or third parties. To answer privacy fears raised by consumers and regulators, telcos must also invest in building digital trust, including actively managing data privacy and having a robust cybersecurity strategy and a framework to guide ethical deployment of AI.

— *Proven use cases and outcomes:* AI-native organizations across industries have deployed AI to achieve four critical outcomes highly relevant to operators across the world: 1) drive revenue protection and growth through personalization, 2) transform the cost structure, 3) enable a frictionless customer experience, and 4) meet new workplace demands. Operators can learn from all of them. Streaming players, for example, have long been known for providing highly curated personalized content recommendations based on past user behavior. To optimize cost and deliver a seamless customer experience, one of the leading US insurance companies leverages AI assistants to reduce and even eliminate human interactions for users to obtain coverage or cancel policies with other carriers. In turn, some of the leading tech companies in the world are known for using AI to highlight the traits of great managers and high-performing teams and use those insights to train company leaders.

— *Technology investments recognized as a business driver:* In a postpandemic world, there is broad consensus among investors and executives that technology investments are not a mere cost center but a fundamental business driver with profound impacts on the bottom line. Despite prospects of economic turmoil and recessionary fears, IT spending is expected to increase by more than 5 percent in 2023, with technology leaders under growing pressure to demonstrate impact on company financials.[2]

— *Operator bets need hypercharging:* As networks and products converge, operators are making bets on becoming cost and efficiency focused, experience-centric, or ecosystem players. AI use cases that are more relevant for each bet can give them a better chance to hypercharge and leapfrog competition.

For the greatest payoff, this shift requires telcos to embrace the concept of the AI-native organization—a structure where the technology

[2] "2023 CIO and Technology Executive Survey," Gartner, October 18, 2022.

is deeply embedded across the fabric of the entire enterprise.

Using AI to reimagine the core business

Telcos have been under relentless pressure over the past decade as traditional growth drivers eroded and economic value increasingly shifted to tech companies. By using AI to its fullest extent, operators can protect their core business from further erosion while improving margins.

As the industry looks to leverage the power of AI, we see six themes gaining prevalence in strategic agendas based on our experience working with telcos across the world.

Hyperpersonalize and architect sales and engagement

Leveraging the breadth and depth of user-level data at their disposal, operators have been increasingly investing in AI-enabled personalization and channel steering.

For example, a hyperpersonalized plan and device recommendation for each line holder could leverage granular behavioral data—such as number of and engagement with apps installed and device feature usage—to create individualized plan recommendations (superior network speed or streaming service add-ons), promos ("Receive unlimited prepaid data to be used for a music streaming service for only $5 per month"), and messaging for specific devices, locations, and events ("Upgrade to the latest device featuring built-in VR"). Subsequently, using audience segmentation tools, customers can be guided to channels that offer an engaging experience while driving the most profitable sales outcome for the telco. A subscriber, for example, with low-digital propensity[3], high ARPU, and high churn risk who is living within a few miles of a store, might be a good candidate to nudge to a device upgrade in-store, leading to better customer experience and potentially stronger loyalty for the operator. Or consider a different scenario: this subscriber uses an advanced 5G network in New York City and is a regular user of fitness apps who travels frequently outside the country. As a result, her telco offers a personalized plan recommendation with superior network access, top fitness app subscription perks, and an attractive international data plan.

Case study: An Asia–Pacific operator that launched a comprehensive customer value management transformation powered by AI (with personalization at the core) achieved a more than 10 percent reduction in customer churn and a 20 percent uptake in cross-sell.

Reimagine proactive service

Earlier investments in digital infrastructure combined with predictive and prescriptive AI capabilities enable operators to develop a personalized service experience based on autonomous resolution and proactive outreach.

With fully autonomous resolution, for example, the system can predict and resolve potential sources of customer dissatisfaction before they are even encountered. After noticing a customer is accruing roaming charges while traveling abroad, the AI system automatically applies the optimal roaming package to her monthly bill to minimize charges. It then follows up with a personalized bill explanation detailing the package optimization and resulting savings for the customer, leading to a surprising and positive CX moment.

Operators are also exploring the redesign of digital service journeys with the help of AI assistants serving as digital concierges. Generative AI technologies, including tools such as ChatGPT, have the potential to enhance existing bots through better understanding of more complex customer intents, more empathetic conversations, and better summarization capabilities (for example, when a bot needs to handover a customer interaction to a human rep). A single unified AI assistant will likely also represent a step change in speed, accuracy, and engagement compared to the interactive voice response systems of today.

[3] Someone who prefers to transact in, and use, assisted channels, such as retail stores and call centers.

An AI-powered service organization is a key ingredient to releasing the full capacity of specialized reps for high-value interactions while improving overall customer experience—one of the key battlegrounds for telcos around the world.

Case study: A leading telco is expected to achieve an approximately 10 percent decrease in device troubleshooting calls, powered by a proactive AI engine that considers the customer's likelihood of calling and issue severity to decide whether to push the most effective resolution via SMS. This proactive engine is also a key element of the operator's ambition to have the highest customer satisfaction scores among competitors.

Build the store of the future

In retail, AI is leading a revolution in the design and running of stores by streamlining operations and elevating the consumer experience.

Some telcos already use virtual retail assistants displayed on floor screens to conduct multiple transactions with customers, including adding balance to a prepaid account and selling prepaid cards and TV subscriptions. A leading European telco leverages AI tools for delivering more-accurate device grading and trade-ins in the store.

The store of the near future includes the following components:

— *Front of house:* Aisle layout and product placement are optimized based on browsing patterns analyzed by machine vision. Digital signage is made relevant to individual customers who are in-store and identified through biometric or geofencing technology. Interactive kiosks serve up personalized promos, service assistance, and wait-time forecasts. Customers are matched with reps who are given nudges with personalized info likely to spark the best interaction and lead to a truly seamless customer experience.

— *Back of house:* Device SKUs are automatically managed to optimize inventory and sales. Stores stock curated assortments based on local preferences surfaced in sales analytics. AI tools such as computer-vision-based grading allows for immediate price guarantees on devices that are traded in.

— *Outside:* Consumers walking near the store receive text or push notifications with a personalized promotion and an invitation to check the product in-store.

*Case study: An Asian telco launched a 5G virtual retail assistant in unmanned pop-up stores. The digital human communicates with customers in a personal and friendly way with engaging facial expressions and body language. She supports customers across multiple transactions, from buying prepaid cards to getting SIM card replacement*s.

Deploy a self-healing, self-optimizing network

The AI-native telco will leverage technology to optimize decision making across the network life cycle stages, from planning and building to running and operating. In the planning and building stages, for example, AI can be used to prioritize site-level capacity investments based on granular data, such as customer-level network experience scores.

In the running and operating phases, AI can prioritize the dispatching of emergency crews based on potential revenue loss or impact on customer experience. AI can also enable a self-healing network, which automatically fixes faults—for example, auto-switching customers from one carrier frequency to another because the former was expected to become clogged. This frees up engineering resources for higher-value-added activities.

Case study: A telecom operator developed an AI-based customer network experience "score" to improve its understanding of how customers perceive their network and to inform network deployment decisions. The AI engine leveraged granular network-level information for every line (e.g., signal strength, throughput) with an ML model to create the score tailored to each customer's individual network experience and expectations. The operator used the score, which directly correlated with impact metrics such as customer churn or network care tickets, to

monitor network performance trending (how the score fluctuated in different regions), to identify opportunities to refine its buildout plan, and to improve how it managed its customer base.

Improve frontline productivity

The AI-native telco also uses AI-enabled tools to optimize workforce planning and coaching of frontline employees across multiple teams, including field force, customer service, and retail associates.

For workforce planning, AI tools enhance traditional applications by forecasting across supply-and-demand metrics for monthly, daily, and intraday time horizons with higher accuracy, more granularity, and full automation. Smart scheduling matches supply with demand, such as reps needed in a call center during particularly busy periods, to meet service-level targets as well as customers' expectations.

Acting as an intelligent coaching manager, an AI-enabled nudge engine provides personalized celebratory and improvement opportunity nudges to employees and their supervisors (Exhibit 1). Coupled with advancements in generative AI, the impact of the AI nudge engine might go even further by, for example, simulating customer responses under different scenarios to train reps.

Case study: A telco operator deployed an AI-enabled scheduling and coaching solution for technicians servicing copper and fiber customers. Resulting efficiency gains included 10 to 20 percent capacity generation and improved customer satisfaction scores.

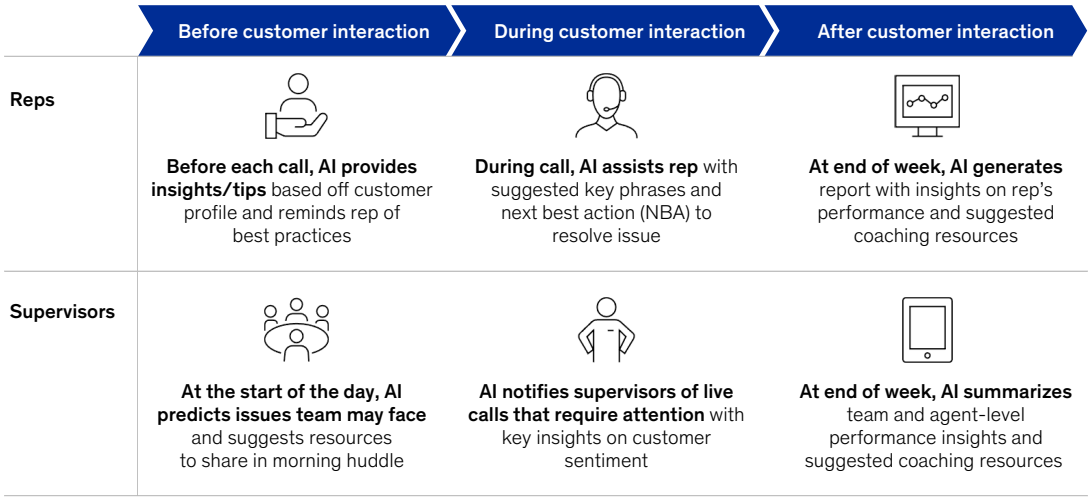
Power intelligent internal operations

AI-powered insights will enhance decision making across business functions, beyond the automation of standardized or low-complexity tasks. In finance, for example, AI can flag outlier invoices for further inspection, while on the accounts receivable side it can predict customers likely to default, triggering mitigating actions. In HR, AI can help flag employees with high attrition or absenteeism risk and the respective drivers while also helping identify informal influencers who can lead change management efforts. Generative AI solutions can help with the development of product marketing copy, the synthesis of customer feedback for research purposes or even enable business users to write simple code to quickly adjust IT applications.

Exhibit 1

The 'AI-native' telco leverages AI to provide tailored coaching recommendations both to reps and supervisors.

Illustrative call-center example



Overall, involving AI in decision making and execution results in higher speed and consistency. Its benefits can be felt everywhere, from contract management and supplier search to onboarding and IT maintenance.

Case study: A UK-based transportation company deployed AI to identify the main drivers of employee attrition and absenteeism. The company then developed targeted interventions for each of the drivers with an estimated 20 to 25 percent reduction in sick pay and attrition costs.

Success factors of AI-native transformation

The *what* of envisioning being AI native is the relatively easier part of this journey; the *how* of making the possibilities reality is the tougher challenge. Working on multiyear projects with operators across the world, we've identified critical best practices in three areas that are the hallmarks of a successful AI-native transformation: building AI, managing it, and driving its adoption.

Building AI best practices

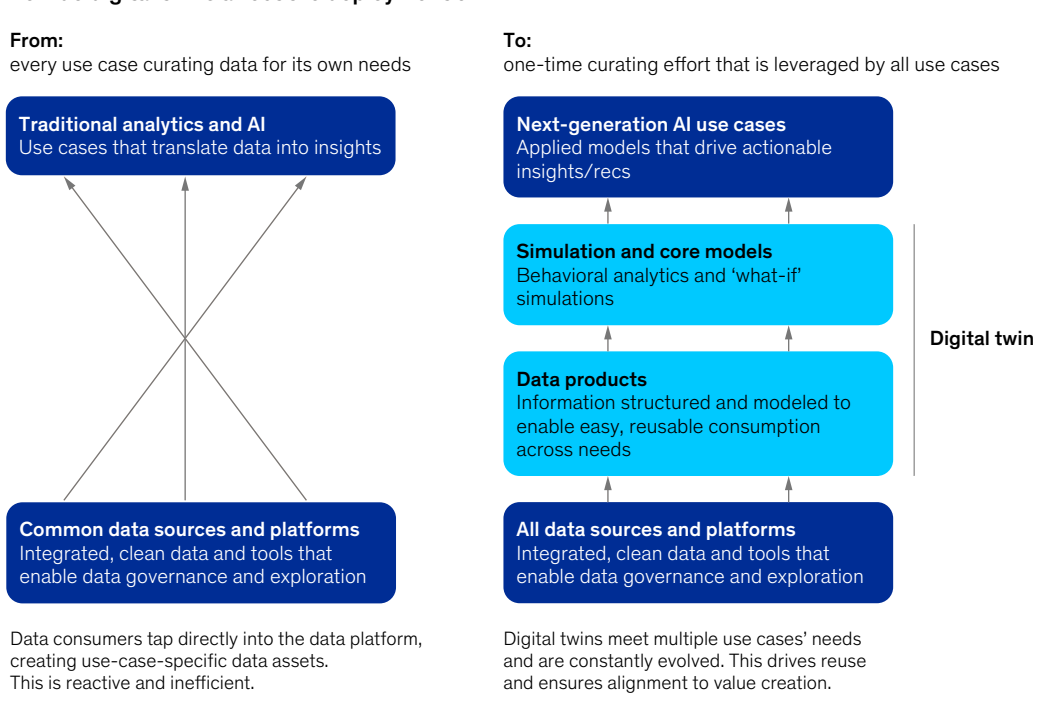
Developing transformative AI requires a carefully calibrated approach that follows these core guidelines:

— Build core AI capabilities in a modular fashion and with reusability in mind, with the potential to be deployed across multiple contexts in the operator. A core forecasting engine, for instance, can be deployed both in a call center and in a retail setting. This will drive higher ROI for AI investments by decreasing time to deploy and preventing duplication of work.

— Tightly integrate AI capabilities with one another based on a model architecture approach that interconnects different AI models to maximize value generation and promote reusability. For example, a digital

Exhibit 2

Digital twins create a single source of truth ('build once, use many times') that speeds up time to market of AI use cases.



propensity model will be built as a core model that becomes an input into multiple customer-facing models.

— Use digital twins as the foundation for all AI. Digital twins—virtual representations of a physical asset, person, or process with a data product at its core—are the key to unlocking reusable AI. The data in a digital twin is intentionally structured and modeled to enable easy, reusable consumption and governance across needs, and to serve as the single source of truth for all models (Exhibit 2).

— Implement machine learning operations (MLOps) best practices to shorten the analytics development life cycle and increase model stability. MLOps typically involve automating the integration and deployment of code underlying AI capabilities.

— Rethink the tech talent strategy holistically. Without a deep bench of engineering talent, an AI-native ambition will remain a mirage. Employers should consider expanding their sourcing net to a wider range of universities and learning environments. It's also critical to improve conditions that developers work under, because developer experience is a top factor in determining an employer's attractiveness.[4] Constraints on which programming languages and cloud providers' tools can be used, for example, can have meaningful impact on a developer's decision to recruit for and stay with an organization, as well as on the developer's productivity. Because tech talent needs are multifaceted, operators should launch a comprehensive list of initiatives across the employee life cycle.

Managing AI best practices

Maintaining and improving AI capabilities depends on an experimental, iterative mindset focused squarely on product and tech innovation.

— Treat AI capabilities as true products by assigning dedicated product managers to oversee them. PMs act as translators between the technical and business teams and are mandated to own the product continuously and develop opportunities to improve it. They ensure that it's never built as a onetime solution.

— Set up AI labs for fast experimentation. Dedicated teams of PMs and data scientists or engineers are granted expedited approval to experiment with new models, test for feasibility, and validate business value before scaling.

— Refresh the AI tech stack at least annually to take advantage of new developments. In recent years, there have been significant enhancements in tooling that drastically transformed AI workflows.

— Speed up IT and data modernization efforts (the complexity of which often slows down AI transformations) by leveraging reference architectures that have been road-tested in multiple transformations across industries. Moreover, build the target cloud-native data architecture following an iterative approach, focused on enhancing the components required for the priority AI use cases first (for example, data streaming might be key to unlock fraud detection use cases).

Driving AI adoption best practices

Taking a comprehensive approach focused on both what goes into and comes out of models is critical for fostering growing usage of AI:

— Ensure AI solutions are considered trustworthy AI, including dimensions such as model explainability, accountability for the outcomes of AI models, and technical robustness.

[4] David Gibson, "New data: What developers look for in future job opportunities," *Stack Overflow*, December 7, 2021.

— Make change management a day one focus. Operators need to involve end users of AI-enabled insights through all the stages of the model development life cycle and invest in formal and informal capability building. Operators will also need to take a hard look at replacing and revamping existing processes as well as management practices and roles to be centered around AI.

Next steps toward building the AI-native telco

In many industries, companies have used AI to make their operations more efficient, drive material enhancements in customer experience, and ultimately used it to bring innovative products and services to market more quickly. Operators can learn from these industries and invest in AI to improve their competitiveness in the coming years of economic uncertainty and competitive turmoil. Many operators have already started to do so.

Organizations that talk about adopting AI but move at a slow pace, hoping that a few innovation projects developed at the fringes of the organization and in silos that will come together to create a snowball effect to holistically change how technology informs business decision making, are likely to fail.

Ultimately, the biggest drivers of AI adoption will be CEO-level sponsorship and full executive alignment throughout the AI-native transformation. The art of the possible with the technology has long surpassed what companies have been able to absorb. Without active support from the top level to proactively address organizational inertia, communicate an engaging change story, model new behavior, promote capability building, and make commitments on the required long-term technological investments, AI-native transformation efforts will not succeed.

The journey to becoming AI native will require operators to create a strategic vision and road map that excites and mobilizes the organization, build priority AI capabilities to gain momentum, and bring everyone together to ensure operating model and change management are set up to drive adoption. Embracing large-scale AI deployment across the organization will follow.

The journey is long and requires commitment, but operators that embrace the path to becoming AI native are more likely to emerge as leaders in the next horizon of transformation.

Joshan Abraham is an associate partner in McKinsey's New York office, where **Miguel Frade** is a consultant and **Tomás Lajous** is a senior partner. **Jorge Amar** is a partner in the Miami office and **Yuval Atsmon** is a senior partner in the London office.

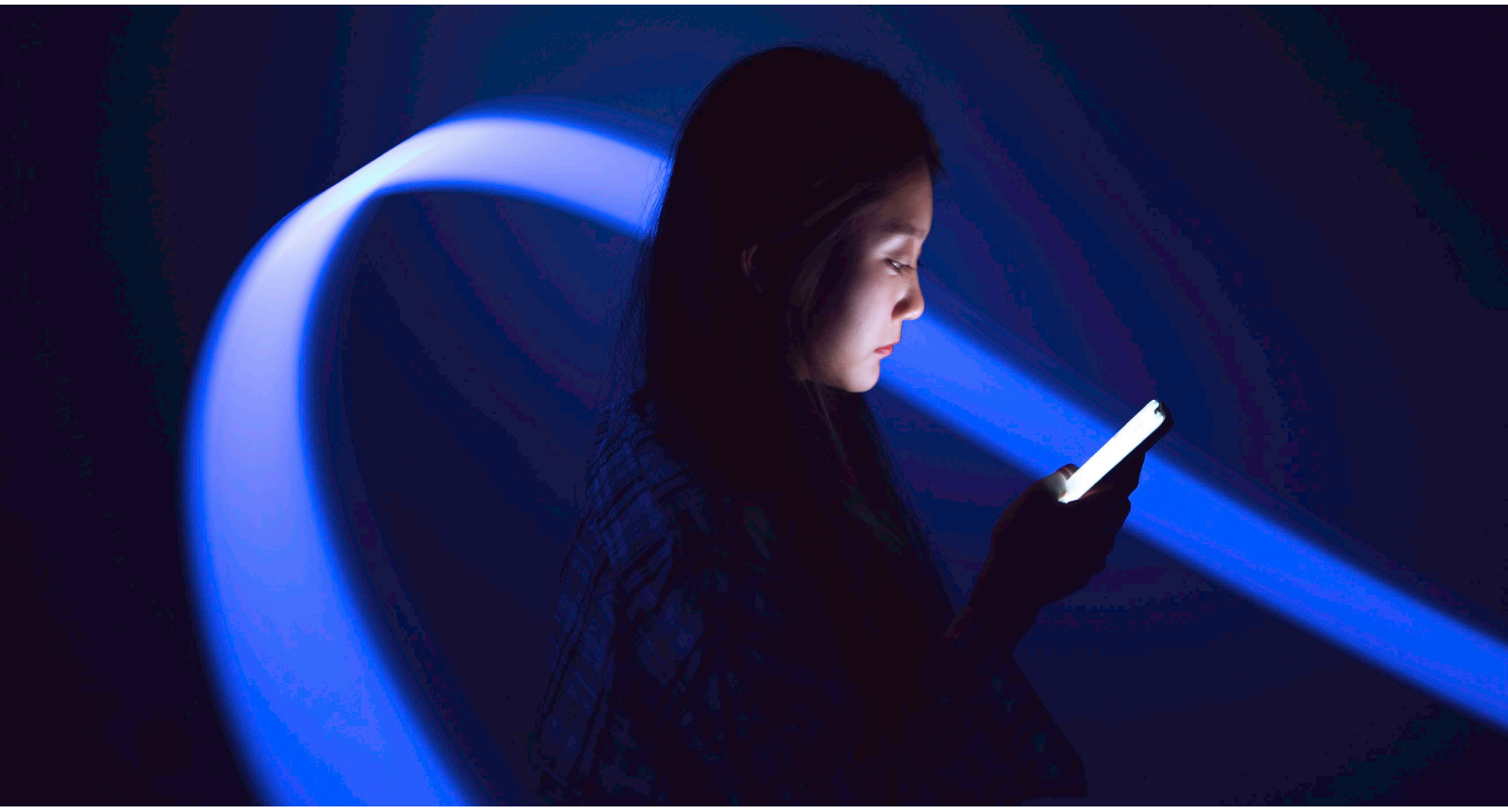
McKinsey & Company

Technology, Media & Telecommunications Practice

How generative AI could revitalize profitability for telcos

The new technology offers the sector a real opportunity to reverse its stagnant fortunes. But seizing it will require embracing innovation and agility to an unprecedented degree.

This article is a collaborative effort by Stephen Creasy, Ignacio Ferrero, Tomás Lajous, Víctor Trigo, and Benjamim Vieira, representing views from McKinsey's Technology, Media & Telecommunications Practice.



February 2024

Amid the intense competition and cost-cutting confronting telcos, early evidence suggests that generative AI (gen AI) could be the catalyst to reignite growth after a decade of stagnation. But will it be a groundbreaking differentiator or simply table stakes?

Already the technology is on track to become a new norm in the industry. Most telco leaders we surveyed[1] say they are developing gen AI solutions that range from pilots to full-scale deployments, and leading telcos such as AT&T, SK Telecom, and Vodafone have made much-publicized early gen AI commitments and launched trials. Some telcos around the world have started to experience significant double-digit percentage impact from this technology. One European telco recently increased conversion rates for marketing campaigns by 40 percent while reducing costs by using gen AI to personalize content. A Latin American telco increased call center agent productivity by 25 percent and improved the quality of its customer experience by enhancing agent skills and knowledge with gen-AI-driven recommendations.

Most impressive is that these telcos deployed the models in just weeks—the first went live in two weeks, and the second in five. For an industry with a mixed track record for capitalizing on new technologies and legacy systems that slow innovation, these early results and deployment times illustrate the potentially transformative power of gen AI.

These aren't one-offs. Pretrained models that can be fine-tuned in days for use cases are readily available, enabling organizations to bring proofs-of-concept to life with minimal up-front investment, achieve impact out of the gate, and scale their efforts. Our experience working with clients indicates the potential for telcos to achieve significant EBITDA impact with gen AI. In some cases, estimates indicate returns on incremental margins increasing 3 to 4 percentage points in two years, and as much as 8 to 10 percentage points in five years, by enhancing customer revenue through improved customer life cycle management and decisively reducing costs across all domains.

However, while nearly all of the 130 telcos we surveyed are doing something with gen AI, our survey findings suggest a palpable sense of caution and skepticism in the industry. More than 85 percent of the executives surveyed are cautious to attribute more than 20 percent revenue or cost savings impact by domain, with the greatest enthusiasm for a radical transformation in customer service (Exhibit 1).

This blend of optimism and restraint highlights the critical juncture the industry faces. Seizing the gen AI opportunity to differentiate services and achieve sustainable growth will require the hidebound industry to embrace innovation, exploration, and agility at an unprecedented level and move from decoupled AI efforts to a holistic, AI-native telco.[2]

The chance for telcos to make this change has never been more accessible. The industry has struggled these last ten-plus years to achieve the potential of "traditional" AI, given the complexity and legacy processes involved. In addition to the significant impact gen AI can bring to bear with entirely new use cases and applications, its ability to learn from vast amounts of diverse data and interact in near-human-like ways may be the tipping point for accelerating broader AI programs and the building blocks that enable them, fueling company-wide transformations.

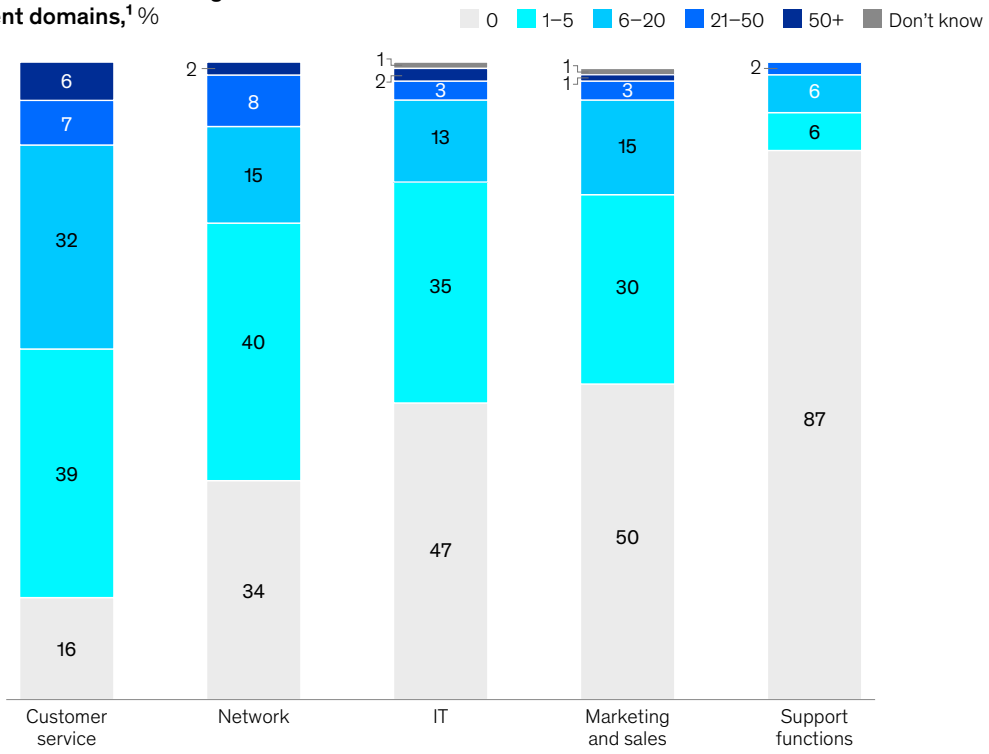
Furthermore, the imperative for such change has never been greater. Because gen AI democratizes access to powerful capabilities, any telco—a small operator or large incumbent—can reshape customer expectations and its organizational efficiency. In doing so, they can potentially narrow previously unassailable competitive advantages and overturn long-standing barriers to growth. Those at the forefront of this movement stand to position themselves to regain growth faster and capture a more significant share of the nearly

[1] The online survey was in the field from November 9, 2023, to December 6, 2023, and garnered responses from 130 telco operators in North America, Latin America, Europe, Africa, Asia, and the Middle East.
[2] See "The AI-native telco: Radical transformation to thrive in turbulent times," McKinsey, Feb. 27, 2023.

Exhibit 1

A large majority of telcos have already cut costs with generative AI use cases in customer service and networks.

Cost reduction attributed to generative AI in different domains,[1] %



[1]Gen AI CxO Survey 2023, n = 130, Q: *What is the impact (% cost reduction) attributed to generative AI in the different domains?* Percentages consider answers only from respondents claiming to have achieved impact and to have at least some use cases in execution; figures may not sum to 100% because of rounding.
Source: McKinsey analysis

McKinsey & Company

$100 billion in incremental value (Exhibit 2). That is in addition to the $140 billion to $180 billion in productivity gains that gen AI will create in the industry above what could be unlocked by traditional AI.

How can telco leaders use the technology to drive AI transformations and unlock new value? What challenges do they face? And what will it take to succeed? This article offers insights into these critical questions, drawing extensively from our research, industry survey, and firsthand experience implementing these technologies.

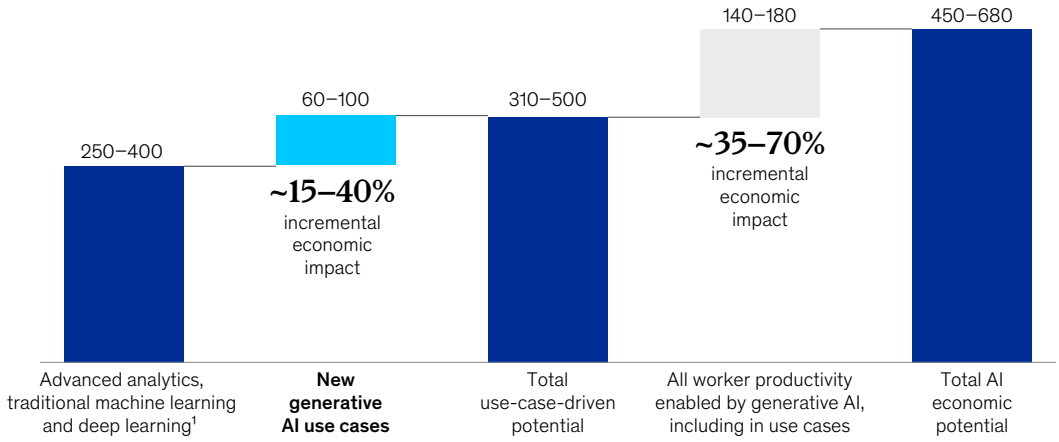
Gen AI today in the telco industry

Gen AI represents the latest advance in AI, and it may arguably be one of the most important. The technology's ability to analyze more and different types of data such as code, images, and text, and to create new content, enables new levels of personalization, performance, and customer engagement. With today's capabilities, many use cases are already possible across network operations, customer service, marketing and sales, IT, and support functions.

Exhibit 2

Generative AI has the potential to unlock value beyond that previously offered by advance analytics and 'traditional' AI.

Generative AI's potential impact on global telecommunications industry,[1] $ billion



[1]Updated use case estimates from "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.
Source: McKinsey Global Institute, *The economic potential of generative AI,* June 14, 2023

McKinsey & Company

These use cases can both enhance existing AI capabilities (through the inclusion of new unstructured data sources) and provide new sources of value (through gen AI and in combination with traditional AI solutions) to deliver significant impact across all key domains. Customer service and marketing and sales currently make up the largest share of total impact (Exhibit 3).

Examples of such use cases based on early pilots include the following:

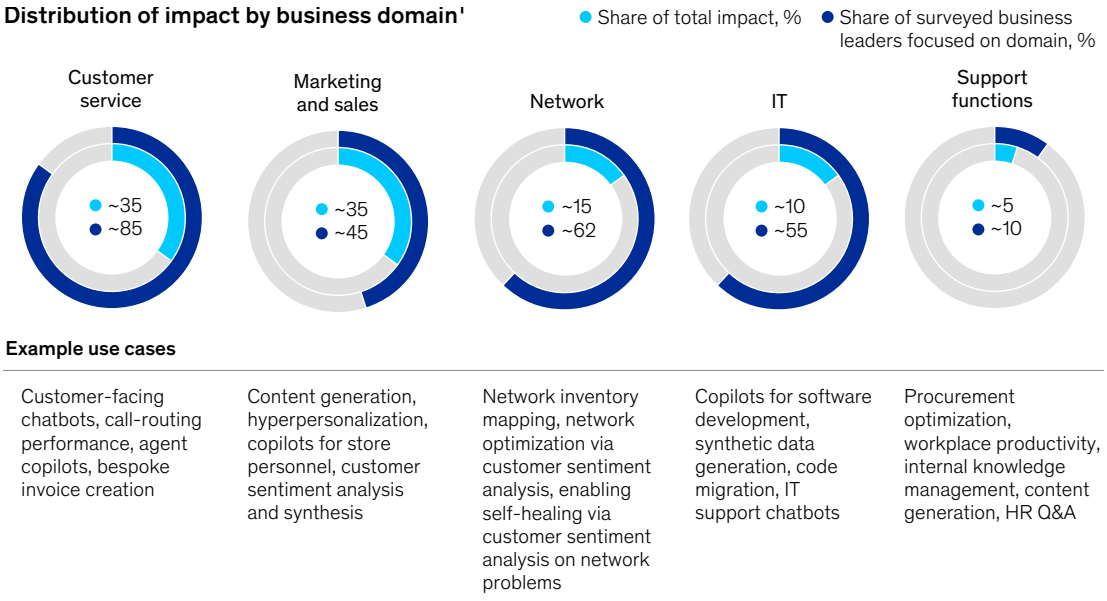
— *In customer service,* where the technology can vastly improve customer experience, increase agent productivity, and enable fully digital interactions. A Latin American telco is enhancing its customer service AI chatbots to improve agent support, a move it anticipates will reduce costs by 15 to 20 percent. The telco also is using gen AI to summarize voice and written client interactions in nearly a dozen use cases, with the expectation that it can reduce associated costs by up to 80 percent.

— *In marketing and sales,* where gen AI enables hyperpersonalization, deeper customer insights, and faster content generation. A European telco is using the technology to identify new sales leads from customer calls, with its pilot project achieving a more than 10 percent conversion rate. The company can now also create personalized messages and visual media to target individual customer microsegments. To do this, the telco feeds a gen AI model standard marketing messages, customer data (including household details, type of phone they use, and where they live), and cognitive biases (for example, whether the customer would be more receptive to messaging that evokes scarcity, such as a limited-time offer, or emphasizes authority, such as endorsements, awards, and industry experience).

— *In network operations,* where gen AI can optimize technology configurations, enhance labor productivity, extract customer insights from social media, and improve inventory and network planning and management through the ability to

Exhibit 3

Gen AI is expected to enable a long list of use cases and deliver significant value to telcos, with customer service and marketing and sales accounting for the largest share of total impact.



[1]The distribution of impact by business domain is based on our experience working with telco companies to deploy gen AI and includes impacts on both capital expenditure and EBITDA.

McKinsey & Company

mine unstructured data. One large telco is using the technology to accelerate network mapping by analyzing and structuring data about network components, including specifications and maintenance information, from supplier contracts. This will enable the telco to more accurately assess component compatibility, maintenance requirements, and more—an effort anticipated to improve operational planning and optimize capital productivity.

— *In IT,* where the technology can accelerate software migrations and development. Gen AI offers telcos a path to reduce their mounting technical debt and enable capabilities previously deferred because of time and resource constraints. Organizations are applying gen AI to streamline the entire software life cycle, from documenting how a new product, feature, or service will be perceived by end users to generating and scanning code for vulnerabilities before launch. One McKinsey study found that software developers can complete coding tasks up to twice as fast with gen AI.[3]

— *In support functions,* where gen AI will reduce the costs associated with back-office operations and improve employee productivity. A European telco uses the technology in a number of ways, including: shortening procurement analysis and negotiation strategy insights from weeks

[3]"Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.

to a few hours, reducing recruiting costs with automated screening and recommendations, improving employee productivity using internal gen AI chatbots and copilots, and automating most internal content generation. Combined, the company anticipates these efforts will improve employee productivity by 30 percent.

New sources of value may also emerge from turning internal uses cases into new products for their customers. For example, a customer care solution may be offered on demand to small business customers seeking ways to improve their own call center's productivity and service.

Telco leaders' view ahead: Real change and real challenges

In the wake of their initial successes, business leaders we surveyed say they plan to maintain or double their budgets for gen AI in the next year and invest in more than 50 dedicated full-time employees to pursue their gen AI ambitions effectively. More than half are already scaling up use cases.

Moreover, survey findings indicate that the technology also had a knock-on effect across all AI initiatives. Compared to responses from McKinsey's 2022 digital twin survey, we see a 30-percentage-point increase in business leaders who want to invest in and focus more on data and analytics.

However, despite the magnitude of the opportunity and the level of interest (and need), our survey found few who follow the kind of holistic approach required to succeed at scale. Only about one-third of telco leaders said they have a capability-building plan for employees on gen AI or are investing in change management efforts—two core building blocks for building a culture of innovation and the test-and-learn mindset. A similar share said gen AI has yet to be treated as an organizational priority, and that proponents of the technology often encounter difficulties in justifying use cases—a clear signal that much of the push has come from the bottom, not the top, and that more work is needed to elevate gen AI to a CEO-led priority. Moreover, finding appropriate talent and obtaining quality data remain significant challenges for telcos, although confidence about solving these rose among surveyed leaders this year as compared to last.

Finally, the survey findings suggest that gen AI has already begun to influence long-standing market dynamics. While European telecom operators have traditionally lagged in AI and technology transformations, survey findings indicate that they are pulling ahead of those in North America in their adoption of gen AI, especially in areas such as network operations (71 percent compared to 58 percent) and IT (67 percent compared to 55 percent). This shift may be a result of greater maturity in managing data privacy. Small and large operators report similar views on where to prioritize, focusing on customer service and IT in similar measure, suggesting the possibility of new competitive pressures emerging for incumbents (Exhibit 4).

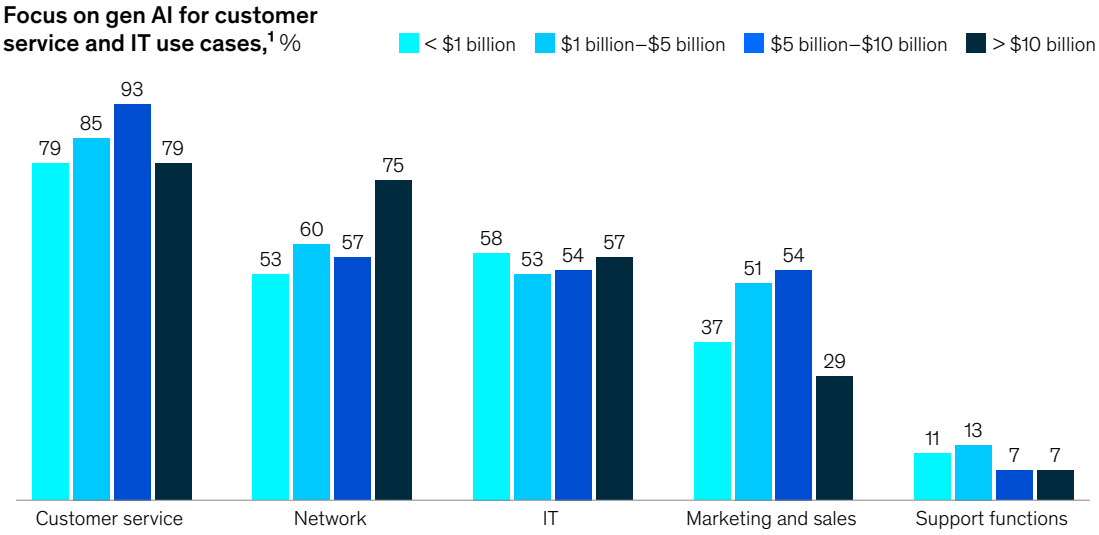
The building blocks for a successful gen AI journey

In order to achieve the above-mentioned impact, organizations will need to move away from the labyrinth of proofs-of-concept and scale the technology. As with any digital or AI initiative, we find there are no shortcuts in doing this. The same core building blocks are necessary, namely (1) a business-led roadmap, (2) the right talent, (3) an operating model for scale, (4) technology built for speed and innovation, (5) quality data that is easily accessible and managed in a responsible and accountable way, and (6) change management to ensure adoption and scaling.[4] These are fundamental pillars in effectively scaling use cases and capturing sustainable impact from gen AI in the journey toward an AI-native telco.

[4] See "Rewired to outcompete," *McKinsey Quarterly*, June 20, 2023.

Exhibit 4

Small and large operators are focused on generative AI use cases for customer service and IT in similar measure.



[1]GenAI CxO Survey 2023, n = 130, Q: *Focus domains in gen AI—Select the dimensions that apply depending on how you define generative AI in your organization.* Percentages may not sum to 100% as this question contained multiple selections.
Source: McKinsey analysis

McKinsey & Company

However, while the same holistic approach is required, gen AI's unique capabilities—its ability to surface new insights from seemingly unrelated data, its reliance on large language models from third-party vendors, and its transformative impact on roles and work—present new challenges that will require greater agility and additional oversight. Next, we outline key differences and provide recommendations on how telcos can best tackle them.

Strategy: Determine when to build, buy, or fine-tune solutions

As with any AI initiative, leaders will need to align on vision, value, and road map, assessing both risks and opportunities, and communicating guidelines for use across the organization. In building a road map, telco leaders will face a choice: use a commercial, off-the-shelf solution if one exists, fine-tune existing large language models with internal data, or build and train a new foundation model from scratch (what we refer to as the "taker," "shaper," and "maker" approach, respectively).

Each is suitable for different use cases and has its own costs, requiring leaders to develop not only a clear vision and strategy for which use cases to pursue, but also how. One mistake we see some telcos making is building common gen AI solutions from scratch—a content generator or call summarization solution for example—when there are nearly a dozen out-of-the-box options on the market today from gen AI start-ups or SaaS vendors injecting gen AI capabilities into existing solutions. Only one-third of surveyed telco leaders say they buy products off the shelf, suggesting that many telcos continue to embrace a do-it-yourself model. This move is likely to slow innovation and distract talent from more differentiating use cases, as it has in the past with other technologies.

Instead, leaders should strongly consider partnering with gen AI solution providers and enterprise software vendors for solutions that aren't very complex or telco specific. This is particularly critical in instances where any delays in implementation will put them at a disadvantage against competitors already leveraging these services. The handful of solutions leaders can concentrate on shaping or making themselves should enable them to differentiate their offerings or address a strategic business priority, such as delivering the best service or network coverage, and drive sustained economic impact. To do this, one CIO at a large telco is bringing together business leaders across all key domains to assess hundreds of potential use cases and build a road map for determining when to build, buy, or fine-tune models, and for prioritizing resources and building momentum with early successes.

Talent: Upskill and expand internal expertise to innovate with gen AI

The speed of innovation that is now possible with gen AI puts new pressure on telcos accustomed to outsourcing tech talent to build in-house AI expertise. Consider the experiences of two telcos—one that continued offshoring and outsourcing tech talent and one that created a dedicated AI team of ten data scientists and engineers. In the time the first telco took to draft requirements for outsourcing gen AI use-case development, the second built and deployed four gen AI solutions.

While this new technology democratizes AI by requiring fewer highly specialized data scientists to build the models, it requires new skills, such as gen AI prompt engineering, which may sometimes be a separate skill embedded within traditional roles. It also requires significantly more data engineers and subject matter experts who understand what data to collect and how, and who can oversee daily quality reviews as new forms of data are generated and consumed by these systems, including user queries, responses, and feedback.

Capturing the full potential will also require significant upskilling of existing staff—everyone from data scientists to business leaders—on gen AI, including the risks of uploading proprietary data into third-party language models. Some telcos are setting up internal certification and university-led training programs to ensure their teams have the right skills and capabilities to innovate and execute with the technology. For instance, a large telco created a badging system to identify gen-AI-ready employees who have completed the company's sessions on use, risk, and effective prompting techniques given by its AI, legal, and risk experts. Following certification, users participate in weekly discussion groups to stay abreast of changes and discuss their successes and challenges. McKinsey research has found that such efforts improve the quality of prompts.

Operating model: Orchestrate efforts enterprise-wide

A significant portion of implemented gen AI solutions can be adapted and reused in multiple use cases. A gen AI chatbot developed to improve agent productivity, for example, can be repurposed with additional fine-tuning or data to answer frequently asked questions by new employees or provide IT support. An off-the-shelf content generation system for drafting sales proposals may also streamline the development of marketing and business plans.

As a result, we're beginning to see telcos adopt more centralized decision making around gen AI development. This shift includes a greater emphasis on adopting reusable services and self-service components, an evolution of key functions, such as risk, FinOps, and transformation offices, to be more focused on gen AI, and the creation of "control towers" that can oversee all gen AI investments and development efforts. In practice, that can mean, for example, prioritizing the use-case pipeline, identifying opportunities for reusability, setting key performance indicators to measure and track impact at the level of both use

case and enterprise, and managing suppliers and risk. A European telco's control tower evaluates the effects of its gen AI transformation based on three dimensions—financial impact, user adoption, and model performance—and aggregates the data in dashboards that enable the company's top executives to keep tabs on the organization's progress. Similarly, a Latin American telco uses a control tower to consolidate and standardize supplier contracts, tracking key metrics such as scope, duration, and renewal to compare providers more easily, identify potential redundancies, and reduce the manual work of digitizing content.

Technology: Create a blueprint for reusability, innovation, and excellence

Organizations will also need a technology blueprint that enables reusability. For instance, the blueprint should include a framework for determining which large language models to use and when (commercial or open-source models, for example, or those that support hybrid workloads). And it should lay out how to scale a pilot—for example, to extend a pilot that serves 100 call agents to serve more than 10,000 agents with the same latency and cost profile. The blueprint should also have a framework for determining which gen AI capabilities can be turned into ready-to-use modules to be plugged into different use cases. One large telco, for example, has begun to identify and develop components designed to fetch product data from a large dataset and generate content from it that could be reused by data science teams across domains such as customer service, network operations, and sales and marketing.

With new gen AI research and capabilities being announced weekly and sometimes daily, technology teams will also need a dedicated gen AI innovation lab to keep abreast of industry changes and test emerging solutions. For example, one large telco's chief data and analytics officer recruited PhD graduates from universities to staff a gen AI innovation lab and build bespoke solutions ahead of the market to gain a competitive edge.

Once new models are deployed, telcos will need to monitor model outputs daily to ensure quality and accuracy do not waver as models learn and adapt their responses based on user queries and feedback. Large language model operations (LLMOps) is an emerging practice that aims to streamline the daily management and monitoring of gen AI models. A key component of LLMOps is a dedicated operations team to oversee all deployed gen AI models, continuously monitoring for issues and rapidly adapting solutions when needed, just as a network operations team might do for network performance. Organizations can start small now and build capability in this area as the field of LLMOps develops. For example, a European telco started by assigning three data scientists to monitor its handful of deployed models and plans to expand the team as more models are deployed.

Data: Capture everything, especially unstructured data, and ensure responsible use

One of gen AI's superpowers is its ability to uncover connections in seemingly unrelated datasets, which has implications for how organizations choose to collect and measure data, and how they manage it to ensure responsible use.

Data collection: Telcos will need to think more broadly about data collection, mapping more data, setting up pipelines for unstructured data, and creating synthetic data to evaluate outputs. A US telco, for instance, has reached beyond its traditional datasets in its work to develop a customer service agent copilot that will reduce average resolution times by 40 percent across more than one million annual chats. As part of their work, the company's data scientists gather institutional knowledge from agent emails and interactions to enable the chatbot to learn from real situations and challenges, and offer detailed descriptions of how to resolve specific issues. The team also creates synthetic data using a large language model to create sample customer questions and answers, with agents reviewing the outputs for accuracy.

Responsible use: These types of more sophisticated data strategies and tactics come with new regulatory, intellectual property, and data privacy concerns. Risks abound in this new era, particularly with customer insights, recommendations, and network optimizations being analyzed and generated by third-party large language models and open-source environments. To address the novel risks, telcos need to expand their data governance programs to address unstructured data. For example, one multinational telco hardware provider created a robust data access process to validate what data can be used in gen AI use cases. Data owners along with legal and security experts work together to validate each use case based on several criteria including the criticality of the data to the business (data that is deemed of high importance cannot be input into commercial large language models); the end users (some users cannot access certain data assets); and the risks if the gen AI solution gives an incorrect answer. The team manages the process in an agile manner using a simple Microsoft Power App to manage and automate the workflow across teams, and conducts monthly forums to review the process and develop improvements. The organization has reviewed more than 200 use cases, rejecting a number due to intellectual property and other risks, to ensure responsible use for the company.

Change management: Ensure adoption and scaling are CEO-led

Every role, including everyone from network technicians to HR professionals, will be impacted by gen AI, making vital the need for leaders to begin preparing their employees now to capture the full value of this transformative technology. With many employees already using the technology in their personal lives, organizations will need to consider how to help them learn to apply the technology in a professional context, upskilling and reskilling staff at scale. Such work can be made easier using gen AI, for example to develop and deliver customized and adaptive training programs, and even to onboard employees.

For example, another European telco saw firsthand the importance of change management and upskilling when it created a gen-AI-driven knowledge "expert" that helped agents get answers to customer questions more quickly. The initial pilot, which didn't include any process changes or employee education, realized just a 5 percent improvement in productivity. As the organization prepared to scale the solution, leaders dedicated 90 percent of the budget to agent training and change management processes, which facilitated the adoption of the solution and resulted in more than 30 percent productivity improvement. The telco also used gen AI to create upskilling programs and provide agents with personalized recommendations for improvement once the solution was rolled out.

Even though so many companies have already achieved real cost savings and revenue improvements with gen AI, these are still the early days of the technology. In the next five years, emerging capabilities—including significant improvements in natural language understanding, advances in human-like reasoning across multiple topics, and availability of real-time solutions with increased accuracy and fewer hallucinations—should unlock even more exciting opportunities beyond the basic improvements seen today.

Combined, these gen AI capabilities will enable telcos to redefine industry standards and set themselves apart in the market. For example, network operations could be enhanced and quality standards radically recast with AI copilots that evaluate images from technicians, provide accurate recommendations for remedies, and automatically initiate interventions or work orders. In sales, cognitive copilots could conduct sentiment analysis on customer calls in real time and guide sales representatives on how best to respond, profoundly altering sales strategies, customer engagement, and overall sales outcomes. Customer service channels

using cognitive chatbots could seamlessly answer complex queries in real time while taking into account privacy and fairness concerns, thereby revolutionizing efficiency while offering customers a human-like experience. Across the enterprise, greater efficiency and productivity could emerge, as domain-specific solutions endowed with an organization's institutional knowledge power an unprecedented wave of automation and AI-driven decision making.

The sudden rise of gen AI has brought the dream of the AI-native telco significantly closer to becoming a reality. With it comes the opportunity for telcos to reverse their recent stagnant fortunes and usher in a new era of growth and innovation. The journey will not be easy, however. To answer the call of gen AI, telcos will need to quickly adopt a culture of innovation and experimentation enabled by the core building blocks shared in this article, one they have previously struggled to build and maintain. With the technology moving so rapidly, those operators that embrace it now are likeliest to create a significant lead that will be difficult for others to follow.

Stephen Creasy is a partner in McKinsey's Copenhagen office, **Ignacio Ferrero** is a partner in the Miami office, **Tomás Lajous** is a senior partner in the New York City office, **Víctor Trigo** is an associate partner in the Madrid office, and **Benjamim Vieira** is a senior partner in the Lisbon office.

The authors wish to thank Joshan Cherian Abraham, Eric Buesing, Michael Chui, Guilherme Cruz, Sebastian Cubela, Andrea Fariña, Roger Roberts, and Kayvaun Rowshankish for their contributions to this article.

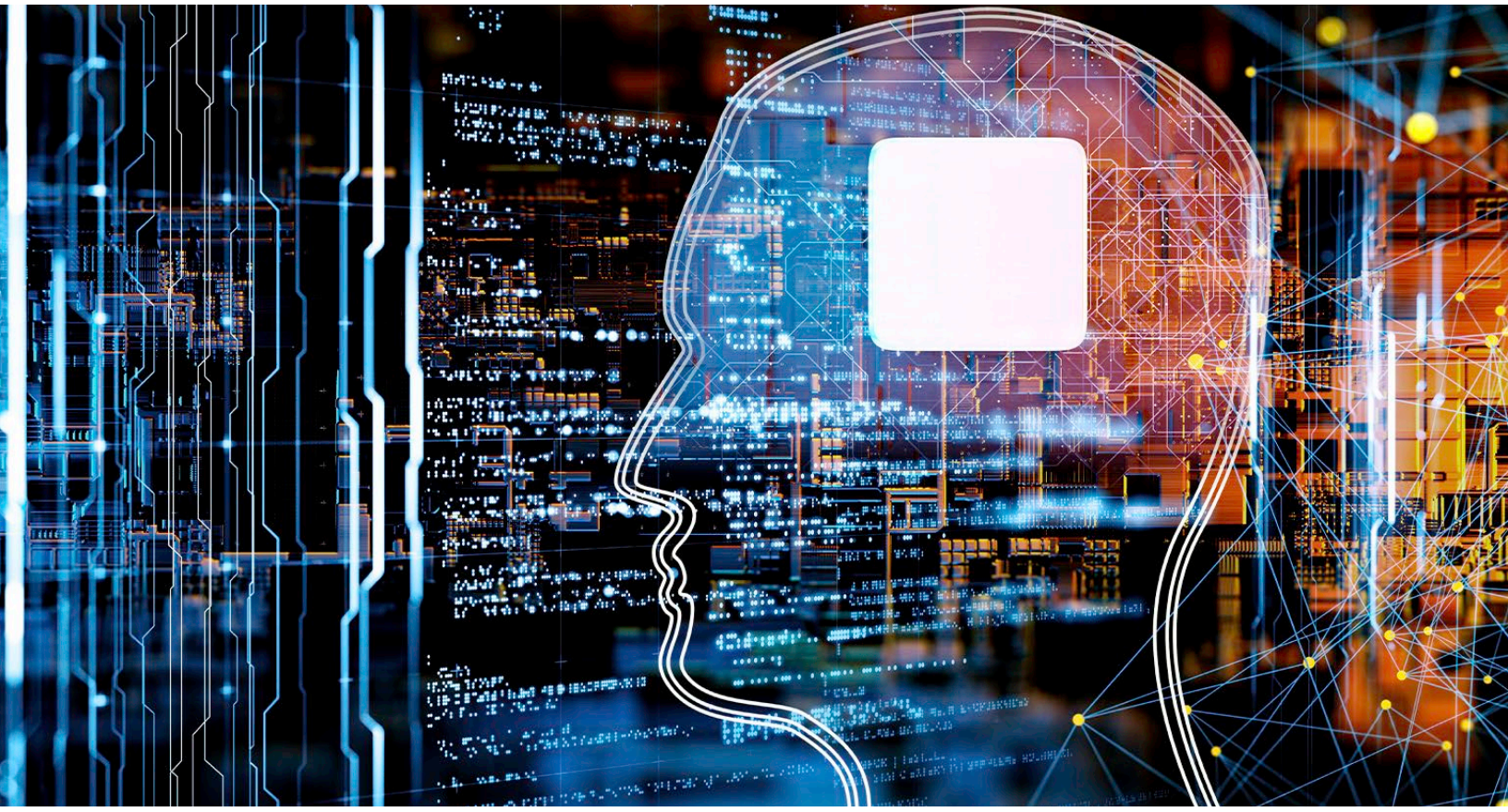
McKinsey & Company

Technology, Media & Telecommunications Practice

Generative AI use cases: A guide to developing the telco of the future

To maximize the opportunity from generative AI, telcos and other TMT players should pursue use cases that have the greatest chance of success and can position the organization for growth and innovation.

This article is a collaborative effort by Ignacio Ferrero, Víctor García de la Torre, Tomás Lajous, Víctor Trigo, and Benjamim Viera, representing views from McKinsey's Technology, Media & Telecommunications Practice.



February 2024

Generative AI's impact on telecom operators is likely to be between $60 billion and $100 billion, according to McKinsey estimates. To capture this new value, organizations will benefit from developing a strategy and implementation road map of use cases across the entire organization. This road map should encompass key areas such as marketing and sales, customer experience (CX), customer service, IT, networks, and support functions. At successful companies, implementation involves a careful choice of priorities and critical decisions based on expected impact and feasibility. Successful companies also tend to favor quick wins to start building capabilities, including the main building blocks: governance, data, talent, and technologies. Strengthening this core enables the move from proofs of concept to use cases at scale.

In the following exhibits, we examine the use cases that hold the most promise for telcos and apply to the broader technology, media, and telecommunications (TMT) sector. Individually, these use cases deliver incremental improvements on existing processes. Taken together, they add up to a radical reimagination of the telco of the future.

This is only the beginning. As gen AI continues to develop and mature, it will unleash even more exciting approaches to fostering creativity, innovation, and growth.

Telco vision: Gen AI will drive the next wave of productivity and innovation

The technology is poised to unlock an array of new efficiencies, improving the way telecom operators attract customers, provide service, maintain networks, and develop software and systems—potentially reigniting growth after a long period of stagnation (Exhibit 1).

Exhibit 1

How generative AI will accelerate the telco of the future.

B2B marketing and sales

5–15% **productivity in marketing functions[1]**

Pre/post-sales activities are AI-driven, with enhanced lead generation via market research and hyper-personalized product offerings and proposals tailored to specific companies.

Negotiation is customized per customer and copiloted with gen AI to enhance customer experience (CX) and boost sales.

Outreach communications (eg, emails and landing pages) are automatically created, leading to cost reduction of ~5–10%.

Support functions

20–30% **automated HR tasks**

Improved productivity across the value chain from speed, automation, and insight creation via tools like HR knowledge bots and enhanced screening of profiles.

B2C marketing and sales

3–5% **sales**

Enhanced CX across the customer life cycle, driven by hyperpersonalization, better digital interactions, and ad hoc product offering and communications (eg, AI-driven content, customized bundles, and contextual promotions/interactions), leading to conversion rate uplift of ~10–15%.

Customer service

Up to 50% **reduction in human-serviced contacts**

Reimagined CX journeys with chatbots able to deliver real-time personalized responses to customer queries, leading to a 30–45% reduction in function costs.

Agents leverage AI-developed scripts and receive instant feedback, accessing relevant customer data for tailored and real-time information delivery.

IT

20–40% **productivity[2]**

Higher productivity as AI-assisted tools accelerate tech delivery timelines with code development copilots (25–30% reduction of time to market), automated documentation, automatic testing, and proactive recommendations.

New frontiers of data, achieved by enhanced value extraction and insight creation from unstructured data (eg, PDFs, social media, audio, images).

Network

15–20% **capital expenditures**

Decision making across all stages of network life cycle, from plan and build to run and operate, optimized by leveraging AI.

Enhanced maintenance with self-healing processes to automatically fix technical issues and boost productivity of engineering resources (eg, a 25–30% reduction in size of incident management team).

[1]Including efficient and effective content creation, product discovery and search personalization, and SEO optimization, as well as reduction in spending on external channels and agencies.
[2]Direct impact of AI on the productivity of software engineering of current annual spending on the function.

Telecom and other TMT players should develop a comprehensive road map of gen AI use cases across the organization

Early movers are building proofs of concept with minimal up-front investment and realizing significant returns (Exhibit 2). In some cases, estimates indicate returns on incremental margins increasing three to four percentage points in two years and as much as eight to ten percentage points in five years.

Exhibit 2

The first step to becoming an AI-native telco player: Identify use cases across the organization.

		Commercial booster		Cost optimization			Key impact, %
Marketing and digital	Marketing	Hyper-personalization of campaigns and communications based on customer profiling		Automatic creation of personalized content for owned and external media at scale			**–60** content generation costs
	Digital		Public facing AI chatbot for sales			Digital campaign ROI optimization	**+10** leads and call-related revenues
Sales and channels	Stores	Boost sales through next best actions and copilot for store employees					**+15** store sales
	Door to door	Agent sale AI doorbot	Perfect sale				**+10** door-to-door sales
	Call center	Commercial front end through perfect sale and copilot for call center agents		Streamlined operations for call center processes		Intelligent policy and scripts enforcement in call center agents across funnels	**–15** costs
Customer care[1]					Front-end efficiencies	Front-end customer experience	**–30** costs
Customer strategy		1:1 customer personalized strategies	Other brands' AI-boosted expansion				**+20** campaign revenue
Network				Planning and deployment		Operations	**–5** network costs
Support	IT			Support function process automation		SQL natural-language prompt generation at scale	**–10** costs
	HR						**–50** outsourcing costs
	Finance	Speech analytics to enhance collections and fraud processes					**–15** reduction in fraud-related costs
Additional areas	Prepaid	Smart campaign management					**+10** revenues
	B2B	Tailored content and proposals	Customer service transformation	Call center enhancement		Data consolidation	**–40** commercial supporting costs
New businesses		Leverage AI in other businesses (eg, energy)					**–5** call-related costs

[1] Includes back office.

McKinsey & Company

Gen AI can boost sales by automatically producing personalized content

Gen-AI-accelerated content creation enables hyperpersonalization across campaigns, enhancing the quality of customer interactions on all channels while improving the effectiveness of top-line efforts and optimizing marketing budget allocation (Exhibit 3).

Exhibit 3

B2C marketing and sales use cases: Bespoke offerings, smarter campaigns.

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Hyper-personalized content	Boost marketing efforts by tailoring communications to specific customer profiles. Gen AI generates text/audio/images and adapts them to specific users (eg, creating more engaging content by adapting language and tone to match user preferences)	Create hyperpersonalized text/email communications (eg, a personalized text to young female with 2 mobile plan subscriptions who lives in the southern US and whose favorite app is TikTok, with key features based on this data)	**15–30% increase in conversion rate (CR)**
Intelligent content generation for marketing campaigns	Automatically create marketing content, including images and videos, for ads, social media posts, landing pages, email campaigns, and other marketing channels	Generate engaging marketing and social media content for new pricing promotions for mobile plans (eg, personalized images for new handset marketing campaign)	**10% increase in CR**
Enhanced global marketing reach through dynamic copy translation	Gen AI's ability to understand context enables automatic translation of marketing copy, supported by quality assurance measures (eg, checking for grammatical errors, clarity, and overall coherence)	Real-time translation of chatbot answers for prepaid international customers and translation of marketing copy in different regions	**5–15% increase in productivity**
Tailored product offerings	Gen AI enhances customer experience with custom product offers, generating new bundles for each customer, based on segment features and historical data	Tailored offer to adults aged 20–30 who consume more gigabytes and fewer call minutes	**5% increase in CR**
Customer feedback analysis	Gen AI enables the extraction of information from unstructured customer feedback sources (eg, call center, web forms, chatbots, etc) to generate insights and identify issues for further analysis and resolution	On the launch of a new technology device, monitor consumer sentiment by automatically analyzing social media posts	**10-point increase in customer satisfaction**

Exhibit 3 continued

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Unlocking best practices of sales champions	Identify specific actions employed by top-performing agents and share them to enhance performance	Upskill sales force (eg, outbound call center sales teams, SME[2] team) by extracting best practices for communication techniques, relationship-building strategies, and persuasive selling approaches	**10–15% increase in CR**
Smart campaign investment optimization	Optimize the investment mix of campaigns deployed, optimizing conversion rates	Improve the budget allocation of lead campaigns across channels	**2–3% increase in leads, calls, and sales**

[1] Not considering cross-effects and interdependencies between different use cases.
[2] Small and medium-size enterprise.

Gen AI can optimize call center efficiencies while enhancing the customer experience

Gen AI helps customers get answers and solutions faster. A major reason is that it supports the work of agents by simplifying their access to relevant data, personalized recommendations, and best practices (Exhibit 4). With this support, agent performance is expected to improve rapidly across the board.

Exhibit 4

Customer service use cases: Smarter agents, happier clients.

Prioritized use case	Future state: What this could look like	TMT application	Potential impact[1]
Proactive root-cause analysis	Leverage advanced analytics to examine call center conversations, identifying recurring issues and root causes to enhance call center operations and improve overall customer experience	Proactively address major reasons for customer care calls, enhancing customer satisfaction (eg, billing issues, roaming coverage)	**30–50% call reduction**
Enhanced agent coaching	Empower agents with improvement opportunities through personalized nudges and hyperpersonalized training content (eg, identifying mistakes in customer interactions) to highlight best practices to enhance the next best action for agent training	Enhance call center agent training to improve performance in future interactions by identifying areas of improvement and provide personalized content (eg, tips to improve script about launch of a new post-paid bundle campaign)	**10–15% cost savings in onboarding and training**

Exhibit 4 continued

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
After-call log-in automation	Gen AI enables the extraction of information from unstructured data (eg, call recordings, post-call notes), automatically extrapolating pertinent information and updating the CRM system	After a customer-agent interaction, such as the canceling of a subscription plan, an automatic ticket is created, along with the identification of root causes and other key attributes	**20–30% increase in typified calls**
Hyper-personalized customer service chatbots	Gen AI enables a 1:1 personalized chatbot experience that proposes an ad hoc solution path, leveraging the customer's existing information and history	Increase the capabilities of the digital sales channel by incorporating upsell, cross-sell, and personalized attention in chatbots, trained with each customer's previous interactions (eg, can be connected to CRM to add additional features)	**10–20% call reduction**
Invoice-focused chatbot	Gen AI is capable of extracting information from invoice documents and providing comprehensive and clear explanations of invoices	Explain increases in price to customers by comparing invoices from last 6 months (eg, customer contracted a plan with more gigabytes)	**5–10% call reduction**
Customer interaction database	Gen-AI-driven database that stores relevant information related to each customer-agent interaction, including summaries of intent, outcome, and resolution path; this can be achieved by leveraging previously untapped data sources, such as call transcripts or audio recordings	Streamline and improve quality of call center agents by facilitating performance evaluations and supporting decision making enabled by easy retrieval and review of call details (eg, classification of customer intent)	**50–60% reduction in after-call work (ACW)**

[1] Not considering cross-effects and interdependencies between different use cases.

Gen AI can boost network operations and reduce call center costs by reducing the number of customers experiencing technical issues

Gen AI's ability to act on unstructured data helps teams optimize network operations and reduce downtimes (Exhibit 5). And stronger processes allow teams to act faster and more effectively.

Exhibit 5

Network use cases: Fewer outages, faster answers, swifter repairs.

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Circuit inventory	Use data from multiple sources to consolidate information and create a unified vision of end-to-end coverage of operator networks and add additional specifications (eg, network coverage, service availability, and technical qualifications)	Create a unified vision of the network's end-to-end coverage (eg, wirelines), leveraging gen AI along with existing unstructured data (eg, supplier contracts)	**2.5–5% reduction in future contract costs**
Chatbot for network data	Deploy an internal chatbot with network information from multiple sources	Assist in organizing the network team's information and responding to other department queries	**10–15% increase in productivity of network employees**
Root-cause identification	Diagnostics and resolution assistant that acts as experts' copilot to enhance capabilities and efficiency in identifying root causes and proposing solutions	In an outage, the system automatically runs a diagnostic analysis to identify root cause, helping network admins resolve the issue quickly	**6–12% reduction in customer care tickets**
Self-healing network	Provide assistance to agents or bots in programming routers for efficient and proactive network management and troubleshooting	In case of a Wi-Fi issue, automatically generate a repair process (eg, reboot network) adapted to contextual data, before dispatching technicians	**20–30% increase in productivity**
Mobile care	Enable proactive actions to prevent and enhance complaint management	Refine the model to identify incidents accurately and integrate with customer care operations	**30–35% decrease in call center calls related to massive incidents**

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Optimization of own vs third-party network usage	Optimize the use of own network by flagging changes or errors in external network parametrization	Add a mechanism for automatic detection of deviations of NRA, RaaS, and VAS traffic between the Comp. C billing platform and the network data	**0.3–0.8% reduction in third-party network consumption**

[1] Not considering cross-effects and interdependencies between different use cases.

Gen AI can be deployed to optimize IT operations and streamline the entire software life cycle

Already, organizations are using gen AI to accelerate software development, testing, and migration—which could free up time and resources for enabling capabilities previously deferred (Exhibit 6).

Exhibit 6

IT use cases: More productive developers, smarter systems.

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Code debugging and optimization	Accelerate tech delivery timelines with automated code development tools enabling low/no-code development (eg, natural-language coding)	Provide support on code review and pull-request completion	**40–55% increase in productivity**
Software development copilot	Accelerate tech delivery timelines with automated code development via gen-AI-driven tools (eg, GitHub Copilot); Copilot offers intelligent code suggestions and autocompletion capabilities and debugging support and documentation	Leverage GitHub Copilot to develop/optimize new code	**25–30% increase in productivity**

Prioritized use case	Future state: What this could look like	Telco/TMT application	Potential impact[1]
Code migration for legacy languages	Interpret, translate, and generate code (eg, migration from legacy systems at scale, automating test development, generating documentation, and performing linting)	Translate COBOL code to Java	**20–30% increase in productivity**
Query assistant	Assist in efficiently and accurately querying (eg, SQL) databases and retrieving specific data, thereby improving data retrieval and analysis processes	Customer value management (CVM) team can independently extract information (eg, new handset sales) to monitor campaign performance	**20–30% increase in productivity**

[1] Not considering cross-effects and interdependencies between different use cases.

Ignacio Ferrero is a partner in McKinsey's Miami office, **Víctor García de la Torre** and **Víctor Trigo** are associate partners in the Madrid office, **Tomás Lajous** is a senior partner in the New York office, and **Benjamim Vieira** is a senior partner in the Lisbon office.

The authors wish to thank José David Vázquez Zelaya for his contributions to this article.

McKinsey & Company

Technology, Media & Telecommunications Practice

Tech talent in transition: Seven technology trends reshaping telcos

With technologies like edge computing, AI, and xRAN transforming telecom, leaders are reassessing how to capture value. Rethinking talent is a key piece of the puzzle.

by Tomás Lajous, Stephanie Madner, Carlo Palermo, and Rens van den Broek

October 2023

The telecom industry is evolving quickly, as businesses and consumers seek out game-changing use cases—from autonomous vehicles to robotic surgery to an unfathomable range of seamless digital interactions—that operate on the back of telcos' substantial 5G infrastructure investments.

Telco leaders are broadly aware of the magnitude of transformation that the moment demands, and many are creating elaborate plans to overhaul everything from business models to operations to customer experience. Ongoing excitement about the potential of AI, driven by advances in generative AI, is pushing the industry to rethink the scope of its transformation plans. However, many telco leaders are struggling to manage the talent implications of these shifts, including determining what talent they need and how to beat out the competition to get it.

The industry has certainly not lacked for engineering PhDs or other markers of technical acumen over the years. But the tech talent market and telcos' position within it have changed dramatically since the generation that is now on the brink of retirement embarked on their careers.

Moreover, not all tech talent is created equal. As telcos evolve to deliver on the opportunities that AI, augmented and virtual reality, and other emerging technologies unlock, they will need to be highly strategic about identifying and attracting talent with the expertise and abilities that each technology demands.

To frame the path ahead, we outline seven broad tech trends that are reshaping the telco industry, along with the talent implications of these trends—including the specific skill sets and capabilities required, as well as those that will likely be phased out. These current tech trends create an urgency for telcos to act now and identify critical talent pools to develop.

We then offer an approach to guide telcos through the complex process of fulfilling their immediate and long-term talent needs. While this approach is rooted in the present trend landscape, it is designed with adaptability in mind and as such will be relevant and applicable to future tech trends that may rise in prominence.

The terrain here is not friendly. Long gone are the days when telcos were the employer of choice for technical talent. Over the next decade, demand for certain tech roles is expected to further increase 20 to 30 percent across US industries—potentially outpacing the supply of recent STEM graduates, which grew just 5 to 10 percent annually from 2015 to 2019. For some roles, telcos' demand is expected to outstrip that of other industries: by 2031, for example, telcos' demand for electrical engineers is expected to grow 24.4 percent, compared with 5.9 percent in other sectors (Exhibit 1).

Telco operators with ambitious goals regarding diversity, equity, and inclusion should be particularly intentional about developing sustainable, long-term talent pipelines. McKinsey research shows that diverse organizations increasingly outperform their nondiverse peers. Telcos' current tech talent pools tend to be less diverse than their overall talent pools; if operators' current talent acquisition and development patterns continue, they stand to become even *less* diverse overall as their tech talent pipeline grows.

Seven tech trends shaping telcos

As digital transformation continues to accelerate, we are on the cusp of further seismic changes to how we work, live, travel, and interact. The seven trends described below are poised to redefine

customers' expectations of telcos—and the role that telcos can play in individuals' lives and the success of organizations.

Each technology will require telcos to grow and stretch in new ways, compelling telco leaders to determine early on where to place bets and to continually refine their priorities as the landscape shifts and technology evolves further. As telcos hire the talent needed to embrace the seven tech trends, they will have less need for skills that can now be automated or are specific to outmoded legacy infrastructure.

At every stage, having the right talent in place will distinguish the leaders from their less successful peers.

1. Ever-expanding connectivity

Fifth-generation (5G) telecommunications infrastructure is dramatically expanding and improving connectivity, and sixth-generation (6G) infrastructure is poised to amplify this trend.

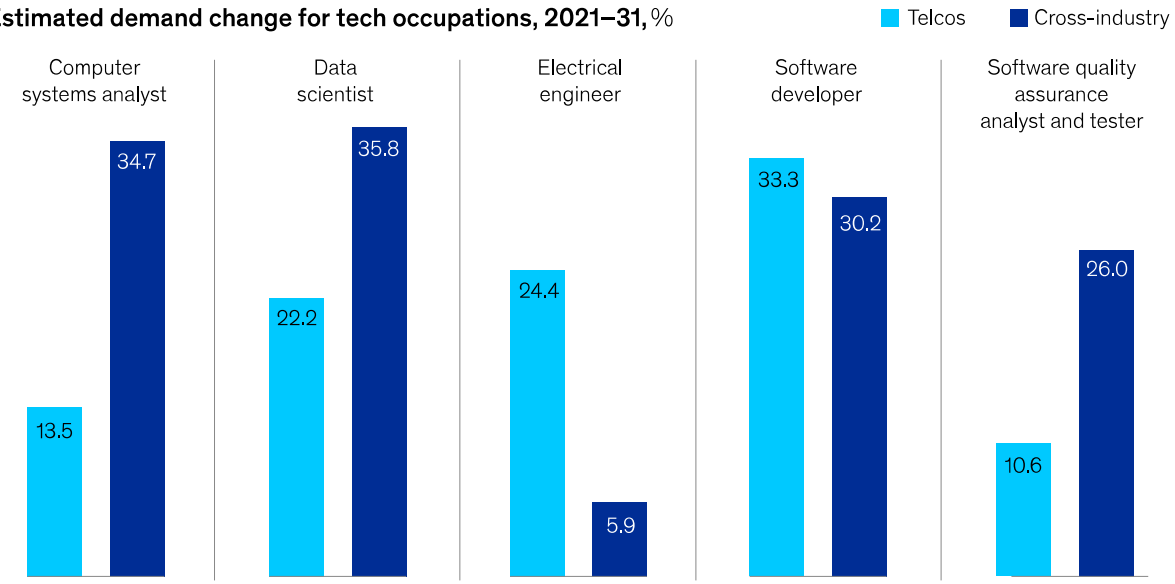
Near-limitless connectivity will pave the way for new services like remote patient monitoring and next-generation customer experiences like virtual dressing rooms and conferences that take place entirely in the metaverse.

Demand for connectivity is expected to increase further as customers seek out these innovative solutions and as the number of connected devices grows to a projected 51.9 billion by 2025,[1] up from 43 billion in 2020. Remote work is also fueling

Exhibit 1

Demand for tech talent critical to telcos is also growing in other industries.

Estimated demand change for tech occupations, 2021–31, %



Source: US Bureau of Labor Statistics 2022

McKinsey & Company

[1] *Worldwide Global DataSphere IoT device and data forecast*, 2021–2025, IDC, July 2021.

demand, with 51 percent of Americans working from home at least one day a week.[2] Moreover, 5G and 6G are estimated to expand connectivity to up to 80 percent of the global population by 2030.[3]

To meet this demand, telcos will need to exponentially increase network capacity, improve data throughput and spectrum, and reduce latency and energy consumption. In addition to extending coverage to individuals, telcos may have an opportunity to raise B2B revenues by developing premium connectivity solutions for specific use cases.

This will require talent with skills in *network and spectrum design* to work on strategy and architecture; *network engineering*, to design architecture and develop applications; *network innovation*, to develop emerging RAN (radio access network), network functions virtualization, Kubernetes, etcetera; *network maintenance monitoring*, to handle emergencies and to fix breaks; and *IoT*, to develop applications, platforms, and APIs.

Engineering and operational capabilities specific to legacy technologies, such as digital subscriber line (DSL), 2G and 3G cellular networks, and traditional cable TV infrastructure, will likely no longer be needed. The talent base, therefore, will need to adapt.

2. Edge computing

As computing workloads are distributed across remote data centers located closer to end users, latency will drop, bandwidth will increase, and organizations will gain more sovereignty over their data. Edge computing allows for real-time data processing, which will unlock use cases across industries—from remote healthcare treatment to remote management of mining operations to sustainability solutions like smart grids that optimize energy consumption.

A recent McKinsey survey of 75 telco executives across North America and Western Europe detected a great deal of interest in a variety of edge computing use cases (Exhibit 2). The survey results showed that a majority of telcos are engaging with edge computing on some level, with a quarter already deploying it or actively planning to scale it. More than half of the executives (55 percent) surveyed said their primary goal is improving network efficiency and performance, while others cited enabling new use cases for businesses (21 percent) or for consumers (18 percent).

As telcos embrace edge computing, they will face rising costs from energy consumption, network maintenance, and investments associated with reconfiguring network backhaul and backbone.

The move toward edge computing requires telco talent with skills in *network and system design* to work on data strategy and architecture; *network engineering*, to enable the installation and integration of devices, software, and systems; *network innovation*, to improve the performance of systems; *network maintenance*, to fix breaks and handle emergencies; *database management*, to manage data storage, distribution, and analysis; and *security*, to minimize fraud, monitor risk, and handle compliance.

The rise of cloud-based solutions, automation, and managed services will reduce demand for *on-site IT setup and maintenance roles*.

3. Next-generation transportation

The first two tech trends, expanded connectivity and edge computing, lay the groundwork for a third: next-generation transportation. The shift toward autonomous, connected, electric, and smart technologies has vast implications for air and land

[2] "Americans are embracing flexible work—and they want more of it," McKinsey, June 23, 2022.
[3] "Connected world: An evolution in connectivity beyond the 5G revolution," McKinsey Global Institute, February 20, 2020.

transportation, with the potential to make human travel and the transport of goods far more efficient and environmentally sustainable.

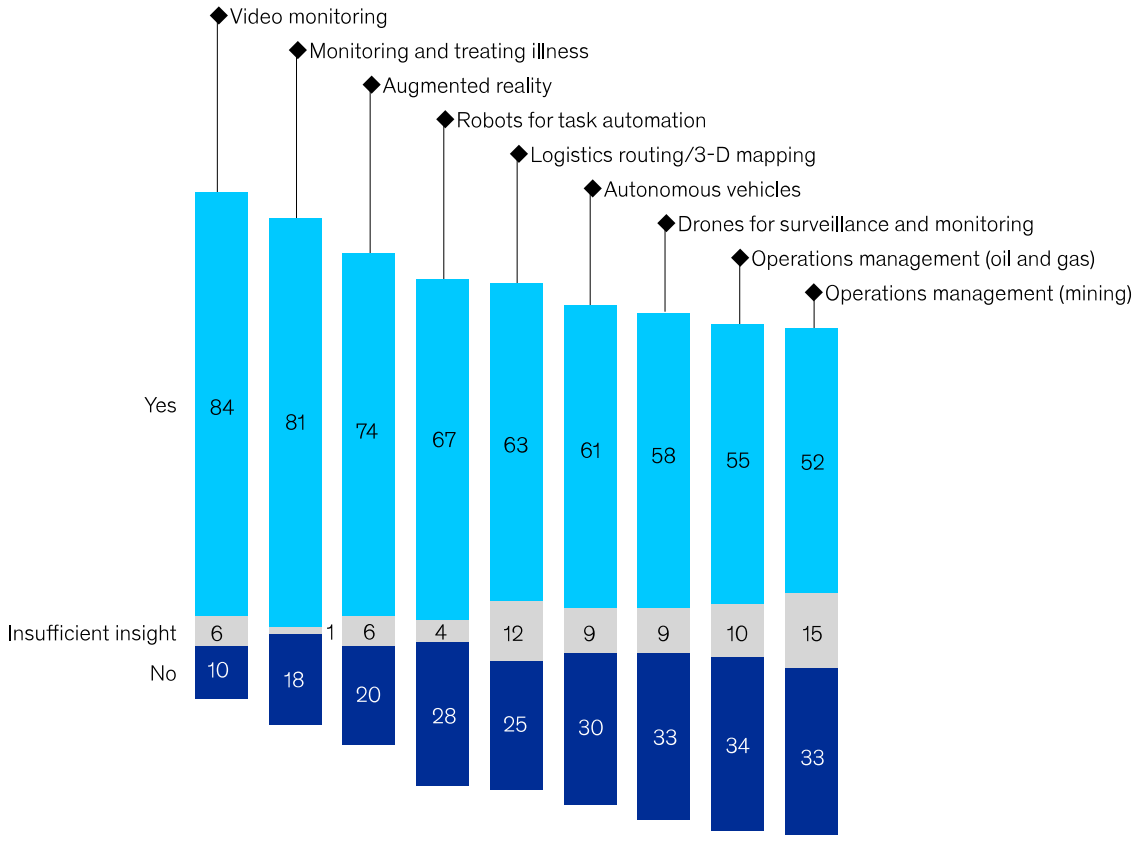
The transportation industry will increasingly prioritize electric, hydrogen-based, and hybrid propulsion as new modes for ground and air mobility. An expected rise in data traffic and autonomous-landing applications may allow businesses to expand their markets, reaching new customer segments in previously unserviceable locations.

As transportation evolves, telcos will need to increase bandwidth for mobility, particularly in remote areas, and provide flawless emergency

Exhibit 2

Most telcos see promise in a wide range of potential edge computing use cases.

Level of telco executives' interest in pursuing different edge computing use cases,
% of telco respondents



Note: Figures may not sum to 100%, because of rounding.
Source: 2022 McKinsey survey of telco executives across North America and Western Europe (n = 75)

McKinsey & Company

backup coverage. They will also have an opportunity to combine core connectivity with vehicular technologies and real-time mobility data to offer solutions like hands-free driving, infotainment, networks of smart electric vehicle chargers, and "vehicle to everything" (V2X) technology—which allows vehicles to connect with their surroundings, including other vehicles and human drivers.

Telcos will need talent with skills in network design to develop algorithms for vehicle connectivity; *network engineering*, *innovation, and maintenance*, to enable vehicle-to-infrastructure connectivity; *automation*, to leverage machine learning and AI for infotainment; *IoT architecture*, to enable voice recognition and gesture control; *UX design*, to enhance the user experience; and *data science*, for collecting and processing data.

4. xRAN

New approaches to RAN can bring flexibility to telcos' relationship with OEMs and even reduce physical-asset requirements like towers, antennas, and cabling, thereby cutting capital and operational spending, accelerating the deployment of new network services, and spurring competition among vendors.

Falling under the umbrella of "xRAN," these new approaches include *open RAN (ORAN)*, which will allow for seamless interoperability among hardware and software from disparate vendors when it reaches maturity; *centralized RAN (CRAN)*, which allows multiple mobile sites to share equipment; and *virtualized RAN (VRAN)*, which supports scalability and network agility by decoupling network hardware from software.

xRAN has the potential to improve telcos' total cost of ownership by allowing them to choose different suppliers for different needs—a dynamic that may encourage new vendors to enter the market and lead to more competitive pricing. This flexibility may lower the risk telcos face when they adopt new hardware or software solutions. And the availability of intelligent, virtualized, and interoperable functions will allow organizations to assemble tailored solutions that increase their capacity.

Our survey of telco executives indicates a strong interest in ORAN in particular, with 76 percent of executives at incumbent telcos and 88 percent of executives at new entrants planning to invest in the new approach. Overall, 60 percent of executives indicated that they plan to use ORAN for at least 20 to 30 percent of new network build-outs.

To accomplish this, telcos will need talent skilled in *agile* working to enhance engineering practices and innovation deployment; *data engineering*, to develop architecture; *cloud*, to develop and test solutions that xRAN enables; *product management*, to enable the evolution of xRAN; and *DevOps*, to build solutions and accelerate the transition to xRAN.

There will be less need for *proprietary hardware knowledge, closed system integration skills, and manual operational capabilities specific to legacy RAN systems*.

5. Trust architecture and digital identity

As organizations build and scale digitally enabled products and services that hinge on collecting vast troves of customer data, trust and privacy will become even more essential. Zero trust architecture, digital identity, and privacy engineering will become more prevalent as companies seek to gain a competitive edge by establishing stakeholders' trust.

To meet rising consumer expectations around digital trust, IT security, and data visibility, telcos

should consider investing in cybersecurity solutions. Those that do so will position themselves to introduce new offerings by building digital identity services on next-generation networks and technologies.

To realize this potential, telcos will need talent with skills in *digital identity development* to provide solutions and trusted technologies; *cybersecurity solution architecture and engineering*, to ensure assessment and secure access to networks and applications; *automation*, to create digital identity solutions and tools; *privacy engineering*, to handle risk and compliance; *network engineering*, to develop apps and architecture; *network maintenance*, to monitor and manage emergencies; and *DevOps*, to automate configuration, continuous delivery, and infrastructure.

Manual production and review of compliance documentation will be phased out.

6. Artificial intelligence

Advances in AI—and in generative AI, in particular—are unlocking opportunities for organizations at every point along the value chain. Telcos can use AI to optimize networks (by managing resources based on real-time traffic and data analysis); proactively address maintenance issues (by analyzing patterns and anomalies to identify problems before they occur); and minimize churn (by analyzing customers' behavior to identify those most likely to leave). By coupling AI-powered cameras and sensors with AI-enabled network maintenance automation, telcos can substantially reduce the costs associated with network infrastructure management.

Generative AI can transform customer experience by supplying customers with highly personalized content, offers, and proactive service-related outreach based on usage patterns, purchase history, and other considerations. By analyzing customer behavior trends, generative AI can enhance product development and accelerate innovation; it might suggest new features for a mobile app or new plans targeting specific customer segments. By using generative AI to simulate sophisticated cyberattacks, operators can identify vulnerabilities and enhance network resilience.

To maximize the AI opportunity, telcos will need talent with skills in *interface design* to create excellent user experiences; *natural language processing engineering*, for AI speech recognition; *data engineering*, to work on data architecture, software, and big data; *data science*, to create mathematical machine learning models; and *security*, to prevent and manage cyberattacks.

As infrastructure is increasingly managed through software, AI will supplant the need for *routine manual troubleshooting*.

7. Quantum technology

Our survey reflects broad consensus among telco executives regarding the impact of quantum technology, with 52 percent saying they believe that quantum will be a differentiating advantage for telcos in the next five years (and an additional 32 percent saying they somewhat agree with this assessment).

Executives see the highest strategic value in developing quantum key distribution (QKD) networks, which allow for the secure exchange of cryptographic keys. Roughly half of executives are already engaging with quantum technology to protect customer data or improve procedures for authenticating users' IoT devices (55 percent), protect telco infrastructure through encryption (53 percent), or encrypt traffic within the network (48 percent).

At the same time, quantum computing will put conventional encryption methods at risk by opening new attack vectors. Organizations are already growing concerned about "harvest now, decrypt later" attacks, in which bad actors steal encrypted data in hopes of using quantum computers to decrypt it in the future. By harnessing quantum technology, telcos can equip themselves with tools to combat these sophisticated threats; QKD, for example, allows communicating parties to be alerted any time an intruder attempts to eavesdrop on an encrypted exchange.

Advances in quantum technology also have the potential to exponentially increase computational performance and the speed of communication. But despite telco leaders' enthusiasm, very few organizations are actively deploying quantum at scale.

To move beyond internal discussions and test-and-learn pilots, telcos will need talent with expertise in *quantum technology (such as quantum algorithms, computer architectures, superconducting circuits, and machine learning); high-performance computing*, to engage with the ecosystem on pilots in areas like QKD; *software and hardware security and crypto-agility*, to prevent cyberattacks and manage cryptography transformations as threat levels and standards evolve; *network engineering*, to design hybrid classical/quantum networks, codevelop and pilot critical equipment for optical communications, and explore the potential of satellite and fiber for quantum communications; and *product management*, to monetize quantum networks and security.

Engineering and operational capabilities specific to traditional network optimization methods, using classical computation, will become less relevant.

Designing a telco talent road map

To get ahead of these seven tech trends, telcos will need to develop long-term strategies for cultivating, attracting, and retaining the right talent, with the right skills. Leading organizations will prioritize diversity at every stage, from strategy design through implementation. They will also engage business leaders from the outset, ensuring that they help shape the tech talent strategy—and that they own it.

Phase one: Identify the talent implications of business strategy

When it comes to incorporating new technologies, one of the most common challenges telcos face is targeting investments directly to those capabilities that fully align with their broader business goals. In the context of a telco's business strategy and competitive landscape, some of the seven tech trends outlined above may be more beneficial or immediately relevant than others.

Telcos can start the process of getting the right tech talent in place by defining their vision for business success over the next three to five years. They can then conduct thorough business impact assessments of the seven tech trends to evaluate how each trend might fuel ambitions like expanding market share, enhancing customer experience, or increasing operational efficiency. Telcos can then work backward to pinpoint the skills and capabilities required to lean into the technologies most pivotal to meeting business objectives. Some examples of the different approaches a telco might take depending on its primary goal include:

— A telco aiming to boost its B2B leadership might opt to invest heavily in edge computing and quantum technology. These tech trends

enable faster data processing and secure communications, which are particularly attractive to business clients.

— A telco aiming to distinguish itself through superior customer experience might prioritize AI, which can enhance customer engagement and customer service through personalized offerings, automated responses, and proactive outreach.

— A telco focusing on global reach and seamless connectivity might prioritize ever-expanding connectivity by investing heavily in 6G. It may deprioritize trends like quantum technology, which may not immediately contribute to expanded network coverage.

— A telco seeking to increase network flexibility while reducing costs might prioritize xRAN. Such a telco may place less emphasis on ever-expanding connectivity, as its main goal would be improving the existing network architecture rather than expanding reach.

— A telco looking to position itself as a technology pioneer might prioritize quantum technology. It may place less importance on trends like xRAN, as its primary focus would be on pushing the boundaries of technology rather than restructuring the existing network.

Phase two: Assess talent gaps and define talent priorities

Once operators are clear on the work that needs to be done and the skills and capabilities required to do it, they can gauge the size of the talent gaps that will need filling and determine which types of talent to prioritize.

They can start by mapping out current hiring and attrition patterns against forward-looking assessments of how demand for each role will change. For operators that are serious about diversifying the workforce, it will be important to understand how demographic variations play out across the talent pool.

At this stage, telcos can also look at broader shifts in supply and demand for different types of roles across the economy. In addition to examining how their hiring rates must evolve, it will be important to consider the sheer numbers involved.

Here are some questions telcos might consider when determining which talent pools to prioritize:

— What is the business value at risk if we don't secure this talent pool?

— How scarce is the market for this type of talent?

— How difficult is it to upskill existing or readily available talent to fill this gap?

— How demographically diverse are the traditional sources for this talent pool?

By clarifying which skills are most critical and which are most easily attainable, operators can focus investments in the areas that matter most. Over time, they can adjust and expand into other talent pools.

Phase three: Design tech talent strategy and operating model

After clarifying their talent needs, telcos can begin formulating a comprehensive talent strategy. This should include a portfolio of innovative initiatives to hire, train, and retain tech talent, as well as outline the necessary infrastructure and other enablers. Key enablers include flexible work arrangements, learning and development platforms that function as hubs for training resources and online courses,

and talent management solutions that span the employee life cycle, from recruitment through succession planning.

When it comes to tech talent, it is difficult to overstate the importance of long-term thinking. Despite recent layoffs at high-profile tech companies, the tech talent shortage persists across industries and could last longer than expected. Tech unemployment in the United States is 2.1 percent—just over half the overall unemployment rate of 3.8 percent.[4] Close to three-quarters of US tech sector workers who were laid off in 2022 found a job within three months, according to data from Revelio Labs.[5] And the demand for tech talent will continue to soar.

Because tech talent pipelines tend to be particularly homogenous, telcos risk regressing on their overall diversity goals if they fail to identify new talent sources. In the United States, just 21 percent of those graduating with a bachelor's degree in computer science are women, 9 percent are Black, and 11 percent are Latino.[6]

By thinking ahead, operators can open up the available solution space and creatively expand their talent pipelines into nontraditional pools and geographical areas. Solutions like these take time but can turn the tide—helping the sector shed its reputation as hierarchical and stodgy, and reposition itself as an agile, nimble, tech-forward employer of choice.

Reimagine career development and the employee value proposition.
As enablers of the most exciting technologies on the horizon, operators have a powerful opportunity to reshape how they are perceived in the talent market. Telcos that pay close attention to tech talent's unique needs, desires, and priorities can reposition themselves to attract the caliber of talent that has seemed hopelessly out of reach for many.

Recent McKinsey research shows that digital talent places a premium on career development and advancement potential—prioritizing these on par with compensation. With new technologies emerging at a dizzying pace, tech workers crave opportunities to learn from experts and peers and to build skills by rotating among different projects and teams.

Our research also found that tech talent values purpose and meaningful work. They want to understand how the tasks that fill their own days support the mission of the broader organization. By creating innovative career development opportunities and a clear sense of purpose, telcos signal that they're attuned to what tech talent wants.

Build holistic university partnerships.
Across industries, leading organizations are reimagining how they work with universities. They are moving beyond discrete internship programs and transactional recruiting efforts that target graduating seniors—instead, they are establishing durable pipelines designed with their specific talent and diversity needs in mind. Done well, these partnerships also provide students with highly sought-after skills and enhance the communities in which telcos operate.

Leading operators are building detailed models to identify the best target universities for such partnerships. These models assess universities' ability to deliver large volumes of high-quality, diverse talent. They also assess the operator's ability to compete with other companies and sectors for talent at each university, based on factors including geographical proximity, alumni presence at the telco, network presence and performance on campus, and ability to meet graduates' salary expectations.

[4] *The tech jobs report*, CompTIA, September 2023.
[5] Hakki Ozdenoren and Devan Rawlings, "You got laid-off. What's next?," Revelio Labs, December 20, 2022.
[6] *STEM jobs see uneven progress in increasing gender, racial, and ethnic diversity*, Pew Research Center, April 1, 2021; *Digest of education statistics*, National Center for Education Statistics.

Holistic university partnerships can take different forms. Qualcomm and several of its top executives or directors have invested heavily over the years in a single university, University of California San Diego; since the late 1990s, company cofounder Irwin Jacobs has invested more than $300 million in the university's engineering programs, healthcare system, and School of Global Policy and Strategy through scholarships and other support for students, faculty, and research.

In another approach, Apple's HBCU C2 initiative, launched in partnership with Tennessee State University, creates coding centers for learners of all ages at historically Black colleges and universities nationwide; it has already expanded into 45 educational institutions. And Lockheed Martin has partnered with the University of Colorado Boulder to fund a research center focused on radio frequency and space systems as well as an engineering management certificate program.

Launch or join tech talent consortiums. Organizations are increasingly seeing the value of partnering with businesses, government agencies, and other players to solve collective talent challenges. Tech talent consortiums, which provide learners of all ages and backgrounds with skill-building opportunities, are promising models for such collaboration. They may be regional or global in nature, and while some focus on specific industries, others take a broader lens to developing cross-industry tech talent.

Telcos may choose to start their own consortium or join one that already exists—like the National GEM Consortium (GEM), which recruits demographically diverse students interested in pursuing graduate-level degrees in applied science and engineering and matches them with member companies in need of their skills. GEM fellows receive stipends and paid summer work experiences with companies including Amazon, Meta, Ford, and Tesla.

As technology continues to evolve, so should telcos' strategies for capturing tech talent. Early movers give themselves the runway to experiment with creative strategies that may pay dividends in the long run—and to adapt and hone these strategies based on rigorous evaluations.

Telcos' future success rests on their ability to make the most of the opportunities that emerging technologies present. Multiple elements will need to fall into place, and telco leaders are developing ambitious transformation plans. But talent strategy is also a critical part of the equation, and it's often not getting the attention it deserves. Business leaders would be well-advised to take the reins in shaping and steering tech talent strategy to ensure they have the people to get the job done.

Tomás Lajous is a senior partner in McKinsey's New York office, where **Stephanie Madner** is an associate partner and **Carlo Palermo** is a consultant; **Rens van den Broek** is a partner in the Bay Area office.

The authors wish to thank Davis Carlin, Vladimir Cernavskis, Zina Cole, Mena Issler, Aaron Kovar, Adam Liang, Kaitlin Noe, Katie Owen, Caterina Priori, and Sirui Wang for their contributions to this article.

3

Deploying Gen AI

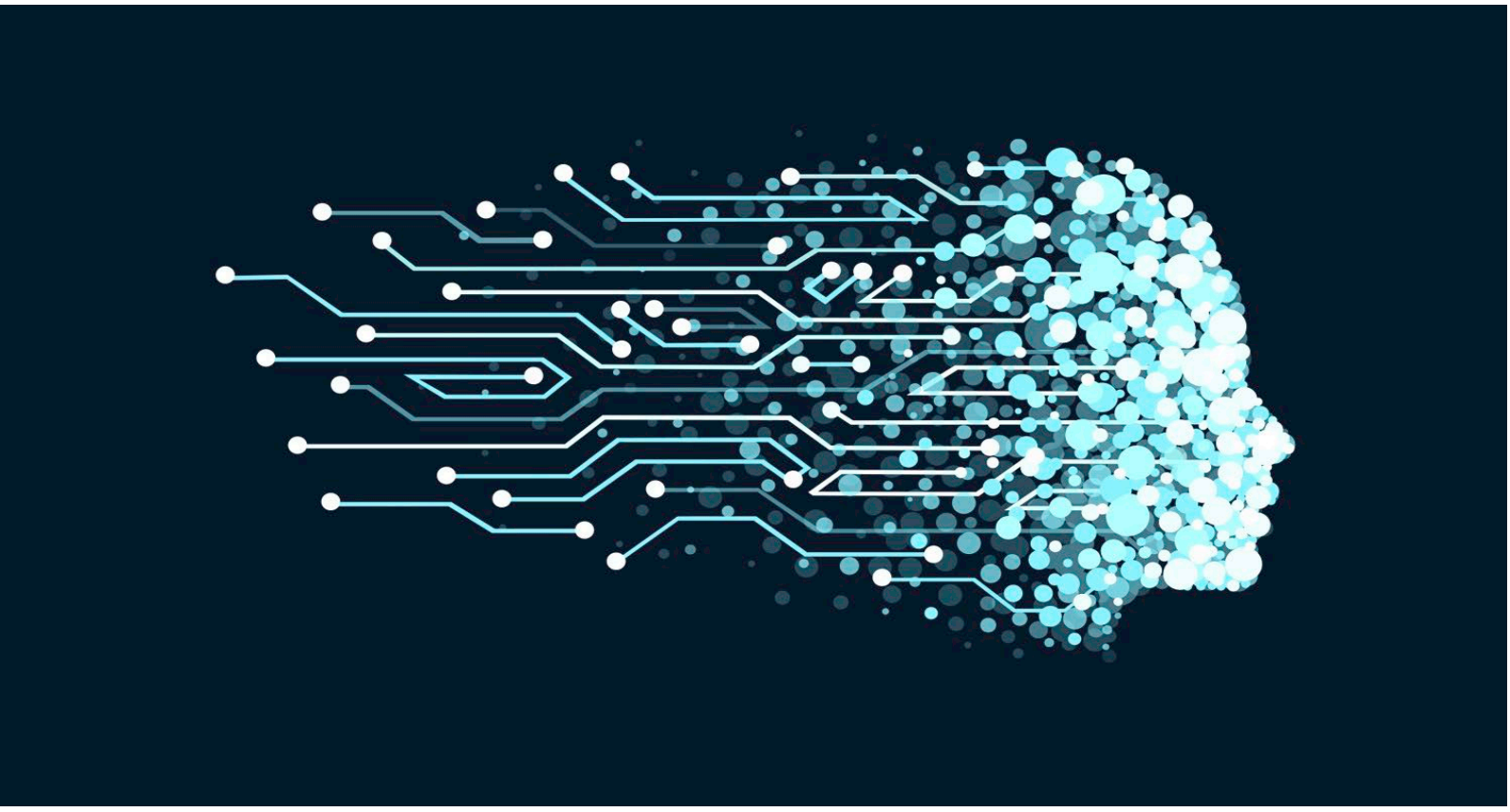
McKinsey & Company

People & Organizational Performance Practice, McKinsey Digital, and QuantumBlack, AI by McKinsey

The organization of the future: Enabled by gen AI, driven by people

Generative AI can empower people—but only if leaders take a broad view of its capabilities and deeply consider its implications for the organization.

by Sandra Durth, Bryan Hancock, Dana Maor, and Alexander Sukharevsky



September 2023

*"**We were behind** on automation and digitization, and we finally closed the gap. We don't want to be left behind again, but we aren't sure how to think about generative AI."*

That's the sentiment shared by many global executives, given the speed with which generative artificial intelligence (gen AI)[1] is advancing in the business world. The technology is accessible, ubiquitous, and promises to have a significant impact on organizations and the economy over the next decade.

Anyone can use gen AI, with little or no formal training or technical know-how. It is being embedded in everyday tools, like email, word processing applications, and meeting software, which means the technology is already positioned to radically transform how people work. And McKinsey research shows that gen AI could enable automation of up to 70 percent of business activities, across almost all occupations, between now and 2030, adding trillions of dollars in value to the global economy.[2]

Meanwhile, technologists keep reminding us that gen AI is only in its nascent stages of development and usage. This smart technology is only going to get more intelligent—and those who don't learn to work with it, starting now, will be left behind.[3]

In this supercharged environment, how can organizations do more than just "keep up"? What strategies, structures, and talent management approaches will business leaders need to adopt to prepare their organizations for a gen-AI-driven future? We examine these and other critical questions in this article.

The situation is evolving rapidly, and there is, frankly, no one right answer to the question of how to successfully roll out gen AI in the organization—business context matters.

But to start, business leaders will need to think broadly about how the rollout of Gen AI could affect their organizations day to day—especially their people. Employees and managers should have a clear understanding of gen AI's strengths and weaknesses and how use of the technology is linked to the organization's strategic objectives. Given the technology's potential to accelerate automation, senior leaders could counter employees' prevailing fears of "replacement and loss" with messaging about gen AI's potential for "augmentation and improvement"—and its ability to significantly enhance the employee experience. Imagine, for example, a world with fewer meetings and more time to think.

The central task for senior leaders, then, is to demystify the technology for others; that will mean taking a step back to assess the strategic implications of gen AI, or the risks and opportunities for industries and business models. As leaders build a compelling narrative for the use of gen AI, they will also need to identify two or three high-impact applications to explore and bring employees along on a value-creating journey—taking gen AI initiatives from pilot test to rapid scaling to "business as usual" status. Senior leaders will also need to commit to building the required roles, skills, and capabilities (now and for the future), so they can continually test and learn with gen AI and stay ahead of competitors.

Are you thinking broadly enough about gen AI's potential impact?

McKinsey research suggests that, because of the emergence of gen AI, about half of today's business activities could be automated a decade earlier than previous estimates had projected.[4] Gen-AI-enabled automation has already begun—and, as the research shows, is likely to affect hours, tasks, and responsibilities for workers across wage rates and

[1] Generative AI is a form of AI that can generate text, images, or other content in response to user prompts. It differs from previous generations of AI, in part, because of the scope of outputs it can create.

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

[3] Paolo Confino and Amber Burton, "A.I. might not replace you, but a person who uses A.I. could," *Fortune*, April 25, 2023.

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

educational backgrounds. In fact, the research shows that gen AI will have an especially profound effect on professions traditionally requiring higher levels of education, such as educators and lawyers.[5]

Gen AI is also likely to inform discussions in the C-suite about how the company creates value and whether the addition of gen AI capabilities allows for industry or business model reinvention. As a result, leaders should ask themselves a range of critical questions relating to the "new" nature of work in gen-AI-enabled organizations, including the following:

What are the organization-wide implications of gen AI? Rather than taking a passive approach to identifying potential use cases and investments associated with gen AI, leaders should view the situation through an "attacker's lens." They should consider all the primary, secondary, and even tertiary effects of gen AI: Which business use cases are highest priority now—and which might be candidates for gen AI enablement in six months, 12 months, and so on? What changes will be required at a functional level to make gen AI enablement possible—for instance, how many more software engineers will the company need? And as gen AI functionality continues to be embedded in common word processing, email, and communications tools (Microsoft's 365 Copilot, for instance), what effect will that have on ways of working across the entire organization? Could gen AI accelerate the shift to a four-day work week? And even more broadly, how might entire industries or business models be fundamentally disrupted?

Does the organization have the right technical talent and risk infrastructure in place? Leaders should consider which operating-model designs will be most effective for ensuring the long-term development of technology talent and the continued evolution of gen AI applications in the organization (see sidebar "Speeding up the search for tech talent"). They should also consider whether that same structure can satisfy the need for gen AI oversight (see sidebar "A powerful resource with potential risks").

How can corporate culture enable or inhibit the adoption and usage of gen AI? Gen AI applications can be the catalyst for culture change—in more ways than one. The applications themselves can create more organizational transparency and connectivity. One company, for instance, is piloting a gen AI application that allows users to ask questions about operations, sales, and other topics, and the tool draws from the company's entire collection of intellectual property to come up with answers that can guide users to the most relevant experts and data. Employees report feeling better informed and more connected. Additionally, the same cultural traits that have been crucial for organizational success during recent economic and business upheavals—such as adaptability, speed, agility, trust, integrity, learning and experimentation, innovation, and a willingness to change—will be even more important if organizations want to become truly enabled by gen AI. To understand why, consider the findings from the 2023 McKinsey Digital survey of 1,000 organizations, which found a significant synergy between organizations with strong, innovative cultures and their ability to increase value through new digital technologies, including gen AI.[6] In previous iterations of that survey, respondents said the biggest obstacle to their digital success was a culture that was averse to risk and experimentation.[7]

How should organizations change their talent management approaches? Gen AI applications will have unprecedented effects on organizations' approaches to talent management. Consider the

[5] "Generative AI and the future of work in America," McKinsey, July 26, 2023.
[6] "Companies with innovative cultures have a big edge with generative AI," McKinsey, August 31, 2023
[7] Reed Doucette and John Parsons, "The importance of talent and culture in tech-enabled transformations," McKinsey, February 20, 2020.

Speeding up the search for tech talent

In the coming months and years, demand for those who have mastered working with and alongside gen AI will skyrocket—especially for those who build and engineer gen AI tools and those who are in the business of generating content via gen AI. (We call the latter "creators," and they can include product managers, marketing managers, and so on.)

To speed up and simplify the search for this critical tech talent amid heavy competition, business leaders should first identify the types of gen AI applications they need to build. They can then use those insights to identify the type and amount of tech talent they will need in the short term—and how to retain that talent for the longer term.

What gen AI applications are we building ourselves? The first decision involves deciding—in collaboration with IT, R&D, and business unit leaders—what applications to build and what applications to adapt from off-the-shelf products.[1] Gen AI applications can be expensive and complicated to build, requiring significant technical know-how. Once built, the applications must be continuously updated or risk losing utility. What's more, training new gen AI applications takes significantly more energy than using or refining existing ones.

Who do we need to build these gen AI applications? Once they know what applications they need to build and buy, senior leaders can examine the technology roles and responsibilities they will need to create value from gen AI. Organizations will need engineering and software development talent, but they will also need translator roles—including implementation coaches, educators, and trainers—to facilitate the understanding and adoption of gen AI across the organization.

How do we develop and retain this tech talent? According to McKinsey research, opportunities for career development, the potential for advancement, and compensation are the top factors technology professionals consider. The chance to learn is another key draw, with professionals reporting a desire to work in an organization that provides employees with opportunities to practice new skills.[2] To meet these requirements and increase the likelihood of retaining top tech talent, senior leaders could explore the use of programs such as peer-to-peer learning, functional rotations that expose technologists to other parts of the organization, and upskilling.[3]

[1] "What every CEO should know about generative AI," McKinsey, May 12, 2023.
[2] "Cracking the code on digital talent," McKinsey, April 20, 2023.
[3] Vincent Bérubé, Dana Maor, Maria Ocampo, and Alex Sukharevsky, "HR rewired: An end-to-end approach to attracting and retaining top tech talent," McKinsey, June 27, 2023.

inevitable impact of gen AI applications on apprenticeship, particularly in the case of knowledge work: imagine a marketing leader uses a gen AI application to write a creative brief that previously would have been developed by a more junior marketing associate. What will happen to the development and mentorship opportunities for both the leader and associate when the learning process is disintermediated by gen AI? What's more, both the content and the delivery of skill-building programs will be affected. A chatbot could guide

A powerful resource with potential risks

Before business leaders can successfully incorporate gen AI into their business strategies and organizations, they must be clear about the risks it may pose and anticipate potential responses; it's the only way to maintain trust with and among employees, investors, and customers.[1]

Among the risks are concerns about the types of biases that may be built into gen AI applications, which could negatively affect specific groups in an organization. There may also be questions about the reliability of gen AI models, which can produce different answers to the same prompts and present "hallucinations" as compelling facts.

Organizations may have trouble shielding some of their intellectual property (copyrights, trademarks, patents, and other legally protected materials) from being inadvertently exposed through a company's gen AI outputs. Similarly, bad actors could plug sensitive customer, supplier, or employee data into a gen AI model to create disinformation, deepfakes, and other types of malicious content.

Organizations will need to take a proactive role in educating regulators about the business uses of gen AI and engaging with standards bodies to ensure a safe and competitive future with the technology.

[1] Jim Boehm, Liz Grennan, Alex Singla, and Kate Smaje, "Why digital trust truly matters," McKinsey, September 12, 2022.

new employees through training on a new technology, at their own pace, on their own terms, allowing them to increase the extent and speed of their learning.[8] Meanwhile, their instructor may use a gen-AI-enabled "teaching assistant" app to create engaging training modules for individuals and groups and to track the progress of both.

These are just a few key organizational considerations; many more are still evolving. Decisions on structure and operating-model design, for instance, will vary from company to company, but whatever the form, our decades-long experience with digital transformations suggests that discussions about value creation must remain at the center.[9] Work processes should enable short, quick cycles of experimentation and iteration and high-quality feedback loops among employees, leaders, and the gen AI applications themselves. To that end, it can be helpful to build small cross-functional teams working end to end on projects and initiatives.

People and gen AI: Building an empowered workforce

Gen AI can be a powerful tool for employee empowerment—even among those who initially perceive it as a threat:

[8] Benjamin S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring." *Educational Researcher*, June–July 1984, Volume 13, Number 66.

[9] Eric Lamarre, Kate Smaje, and Rodney Zemmel, *Rewired: The McKinsey Guide to Outcompeting in the Age of Digital and AI*, New York, NY: Wiley, 2023.

It can augment the employee experience. Gen AI applications can assist employees in ways that many workers may not even expect. For instance, gen AI can suggest the new lines of code required to update a financial-reporting system or outline the A and B versions of a marketing campaign or otherwise create first drafts that human employees can take and implement in live production environments. And by facilitating the training and upskilling process, gen AI applications can help employees pick up new skills more quickly. A recent study, for instance, found that software engineers completed their coding tasks up to twice as fast when using gen AI and reported more satisfaction with the process.[10]

It can empower middle managers. The benefits of gen AI can accrue not just to frontline workers but also to middle managers. In fact, as the people closest to employees, middle managers have a critical role to play in increasing employees' comfort with both short-term gen-AI-enabled work and long-term collaborations with the technology.[11] And as their own direct reports learn to work with gen AI, middle managers may find themselves overseeing more and different kinds of work streams, moving at a pace never seen before. At the same time, the use of gen AI can free up more capacity for middle managers, so they can shift their attention to higher-value leadership tasks, such as strategy-focused work and people management.

It can help organizations reinvent their talent management practices. The emergence of gen AI presents an opportunity for organizations to hone their approaches to attracting, retaining, and developing talent—particularly when it comes to creators and tech professionals. HR professionals could use gen AI to send personalized outreach emails to candidates and to design job search experiences for candidates in underrepresented groups; research suggests this work could dramatically increase the number and diversity of applications for various roles.[12] Gen AI applications could also help companies match new hires with mentors and coaches to improve the onboarding experience, upskill talent, and streamline administrative tasks.

It can prompt senior leaders to lead differently. Senior leaders face the dual responsibility of quickly implementing gen AI today and anticipating future versions of gen AI technologies and their implications. More than anyone else in the organization, they will need to be evangelists for gen AI, encouraging the development and adoption of the technology organization wide. That will mean working with other business unit and technology leaders to allocate resources to update technology infrastructure and take any interim process steps required to facilitate the gen AI rollout—for instance, moving applications to private cloud-hosted environments. In fact, a central task for senior leaders will be to find ways to forge stronger connections between technology leaders and the business units. One company, for example, launched a Slack channel devoted to ongoing discussion of gen AI pilots. Through such forums, employees, product developers, and other business and technology leaders can share stories about their experiences with gen AI, whether and how their daily tasks have changed, and their thoughts on the gen AI journey so far.

As they would when introducing any new technology, senior leaders should speak clearly about the business objectives of gen AI, communicating early and often about gen AI's role in "augmenting versus replacing" jobs. They should paint a compelling picture of how various aspects of the organization will be rewired through gen AI—technically, financially, culturally, and so on.

[10] "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.

[11] *People & Organization Blog*, blog post by Emily Field, Bryan Hancock, Ruth Imose, and Lareina Yee, "Middle managers hold the key to unlock generative AI," McKinsey, July 19, 2023.

[12] Justin Friesen, Danielle Gaucher, and Aaron C. Kay. "Evidence that gendered wording in job advertisements exists and sustains gender inequality," *Journal of Personality and Social Psychology*, 2011, Volume 101, Number 1.

Of course, if senior leaders don't understand the technology themselves, it will be more difficult to make this case for, and lead their teams into, a gen-AI-enabled future. One way for leaders to stay plugged in is to establish forums that provide ongoing professional education on advances in AI technology and applications. Another approach is to carve out time during planning meetings to consider forward-looking questions such as, "Is our approach to gen AI today flexible enough to accommodate the next iteration, and the one after that?" and "Which process steps or roles will we be able to reinvent with the *next* iteration of gen AI?"

Time to flex your gen AI muscle

Although generative AI burst onto the scene seemingly overnight, CEOs and other business leaders can ill afford to take an overly cautious approach to introducing it in their organizations. If ever a business opportunity demanded a bias for action, this is it. By taking the following three steps simultaneously, and with a sense of urgency, leaders can do more than just "keep up"—they can capture early gains and stay ahead of competitors.

Demystify gen AI for everyone. Senior leaders themselves should develop a deep understanding of gen AI and associated capabilities themselves so they can help to demystify the technology for the rest of the organization. They can then help to introduce mechanisms for managing uncertainties about gen AI where they exist—for instance, establishing clear guidance regarding the use of gen AI tools in hiring and recruiting where AI model biases could emerge.

Identify two or three high-impact use cases—and just get started. Senior leaders should carefully consider their investments in gen AI pilots, and "go big" on those that show the greatest promise of scalability and long-term value—whether it's an application that simplifies financial reporting or one that enhances onboarding for new hires. As part of this vetting process, senior leaders should consider the business or industry risks or opportunities associated with implementing the gen AI pilot, as well as how hard or easy it will be to move the pilot into production and make it a part of employees' day-to-day experiences. Once that vetting has happened, senior leaders should steer resources accordingly and take care to monitor and measure the outputs from gen AI initiatives and pilots. Remember, some gen AI initiatives may show impact in the next 12 months, while others may require investment now to yield results in two to five years. The longer-term goal, then, should be to set up a sustainable engine for the rapid upskilling of employees and scaling of gen AI and other digital capabilities.

Commit to building the necessary roles, skills, and capabilities—now and in the future. Senior leaders should commit to building employees' gen AI skills so they can use the technology judiciously and successfully in their day-to-day work. It's not a one-and-done process; leaders will need to continually assess how and when tasks are performed, who is performing them, how long tasks typically take, and how critical different tasks are. Through this process, leaders can better understand current and future talent needs and determine how best to redeploy and upskill talent. Indeed, upskilling programs will take on greater importance than ever, as employees will need to learn to manage and work with gen AI tools that are themselves ever evolving. Leaders should also keep in mind that gen AI itself may facilitate the creation of content for, and automated or personalized delivery of, such upskilling programs.

In the time it took to read this article, gen AI applications have already gotten that much smarter. Leaders can put that intelligence to good use. It's clear that much of the value of gen AI will come from tailoring it to organization-specific use cases—but the successful integration of gen AI requires experimentation and iteration. There is no time to sit back and learn from others' mistakes. Invest deliberately. Get your hands dirty. Start now.

Sandra Durth is an associate partner in McKinsey's Cologne office, **Bryan Hancock** is a partner in the Washington, DC, office, **Dana Maor** is a senior partner in the Tel Aviv office, and **Alexander Sukharevsky** is a senior partner in the London office.

The authors wish to thank Jan Bouly, Michael Chui, Neel Gandhi, Randy Lim, Federico Marafante, Maria Ocampo, Joachim Talloen, Alon Van Dam, and Anna Wiesinger for their contributions to this article.

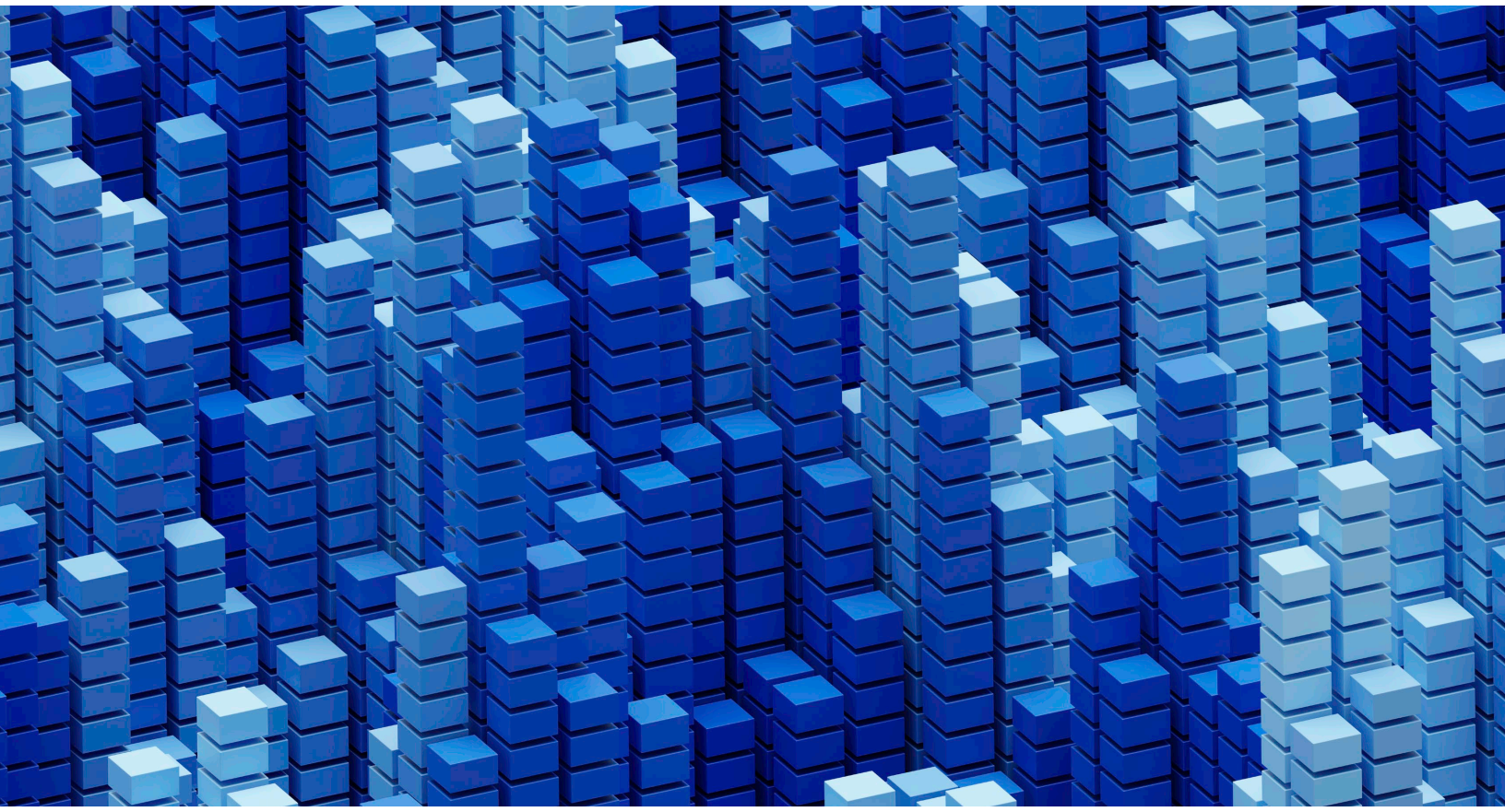
This article was edited by Roberta Fusaro, an editorial director in the Waltham, Massachusetts, office.

McKinsey Digital

The data dividend: Fueling generative AI

Data leaders should consider seven actions to enable companies to scale their generative AI ambitions.

This article is a collaborative effort by Joe Caserta, Holger Harreis, Kayvaun Rowshankish, Nikhil Srinidhi, and Asin Tavakoli, representing views from McKinsey Digital.



September 2023

If your data isn't ready for generative AI, your business isn't ready for generative AI.

Our latest research estimates that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion in annual economic benefits across 63 use cases.[1] Pull the thread on each of these cases, and it will lead back to data. Your data and its underlying foundations are the determining factors to what's possible with generative AI.

That's a sobering proposition for most chief data officers (CDOs), especially when 72 percent of leading organizations note that managing data is already one of the top challenges preventing them from scaling AI use cases.[2] The challenge for today's CDOs and data leaders is to focus on the changes that can enable generative AI to generate the greatest value for the business.

The landscape is still rapidly shifting, and there are few certain answers. But in our work with more than a dozen clients on large generative AI data programs, discussions with about 25 data leaders at major companies, and our own experiments in reconfiguring data to power generative AI solutions, we have identified seven actions that data leaders should consider as they move from experimentation to scale:

1. **Let value be your guide.** CDOs need to be clear about where the value is and what data is needed to deliver it.

2. **Build specific capabilities into the data architecture to support the broadest set of use cases.** Build relevant capabilities (such as vector databases and data pre- and post-processing pipelines) into the existing data architecture, particularly in support of unstructured data.

3. **Focus on key points of the data life cycle to ensure high quality.** Develop multiple interventions—both human and automated—into the data life cycle from source to consumption to ensure the quality of all material data, including unstructured data.

4. **Protect your sensitive data, and be ready to move quickly as regulations emerge.** Focus on securing the enterprise's proprietary data and protecting personal information while actively monitoring a fluid regulatory environment.

5. **Build up data engineering talent.** Focus on finding the handful of people who are critical to implementing your data program, with a shift toward more data engineers and fewer data scientists.

6. **Use generative AI to help you manage your own data.** Generative AI can accelerate existing tasks and improve how they're done along the entire data value chain, from data engineering to data governance and data analysis.

7. **Track rigorously and intervene quickly.** Invest in performance and financial measurement, and closely monitor implementations to continuously improve data performance.

1. Let value be your guide

In determining a data strategy for generative AI, CDOs might consider adapting a quote from President John F. Kennedy: "Ask not what your business can do for generative AI; ask what generative AI can do for your business." Focus on value is a long-standing principle, but CDOs must particularly rely on it to counterbalance the pressure to "do something" with generative AI. To provide this focus on value, CDOs will need to develop a clear view of the data implications of the business's overall approach to generative AI, which will play out across three archetypes:

— **Taker:** a business that consumes preexisting services through basic interfaces such as APIs. In this case, the CDO will need to focus on making quality data available for generative AI models and subsequently validating the outputs.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[2] McKinsey Data & AI Summit 2022.

— **Shaper:** a business that accesses models and fine-tunes them on its own data. The CDO will need to assess how the business's data management needs to evolve and what changes to the data architecture are needed to enable the desired outputs.

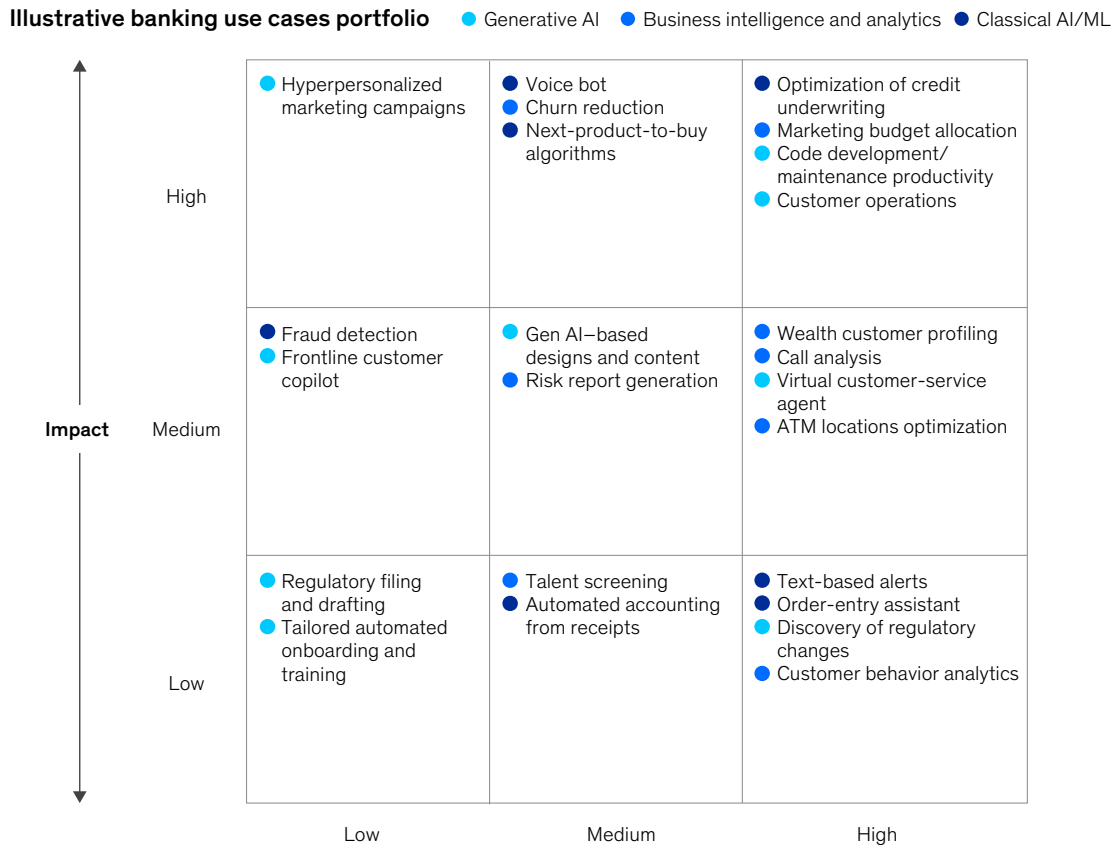
— **Maker:** a business that builds its own foundational models. The CDO will need to develop a sophisticated data labeling and tagging strategy, as well as make more significant investments.

The CDO has the biggest role to play in supporting the Shaper approach, since the Maker approach is currently limited to only those large companies willing to make major investments and the Taker approach essentially accesses commoditized capabilities. One key function in driving the Shaper approach is communicating the trade-offs needed to deliver on specific use cases and highlighting those that are most feasible. While hyperpersonalization, for example, is a promising generative AI use case, it requires clean customer data, strong guardrails for data protection, and pipelines to access multiple data sources. The CDO should also prioritize initiatives that can provide the broadest benefits to the business, rather than simply support individual use cases.

As CDOs help shape the business's approach to generative AI, it will be important to take a broad view on value. As promising as generative AI is, it's just one part of the broader data portfolio (Exhibit 1). Much of the potential value to a business comes from traditional AI, business intelligence, and machine learning (ML). If CDOs find themselves spending

Exhibit 1

Take a portfolio view on value.



90 percent of their time on initiatives related to generative AI, that's a red flag.

2. Build specific capabilities into the data architecture to support the broadest set of use cases

The big change when it comes to data is that the scope of value has gotten much bigger because of generative AI's ability to work with unstructured data, such as chats, videos, and code. This represents a significant shift because data organizations have traditionally had capabilities to work with only structured data, such as data in tables. Capturing this value doesn't require a rebuild of the data architecture, but the CDO who wants to move beyond the basic Taker archetype will need to focus on two clear priorities.

The first is to fix the data architecture's foundations. While this might sound like old news, the cracks in the system a business could get away with before will become big problems with generative AI. Many of the advantages of generative AI will simply not be possible without a strong data foundation. To determine the elements of the data architecture on which to focus, the CDO is best served by identifying the fixes that provide the greatest benefit to the widest range of use cases, such as data-handling protocols for personally identifiable information (PII), since any customer-specific generative AI use case will need that capability.

The second priority is to determine which upgrades to the data architecture are needed to fulfill the requirements of high-value use cases. The key issue here is how to cost-effectively manage and scale the data and information integrations that power generative AI use cases. If they are not properly managed, there is a significant risk of overstressing the system with massive data compute activities, or of teams doing one-off integrations, which increase complexity and technical debt. These issues are further complicated by the business's cloud profile, which means CDOs must work closely with IT leadership to determine compute, networking, and service use costs.

In general, the CDO will need to prioritize the implementation of five key components of the data architecture as part of the enterprise tech stack (Exhibit 2):

- **Unstructured data stores:** Large language models (LLMs) primarily work with unstructured data for most use cases. Data leaders will need to map out all unstructured data sources and establish metadata tagging standards so models can process the data and teams can find the data they need. CDOs will need to further upgrade the quality of data pipelines and establish standards for transparency so that it's easy to track the source of an issue to the right data source.

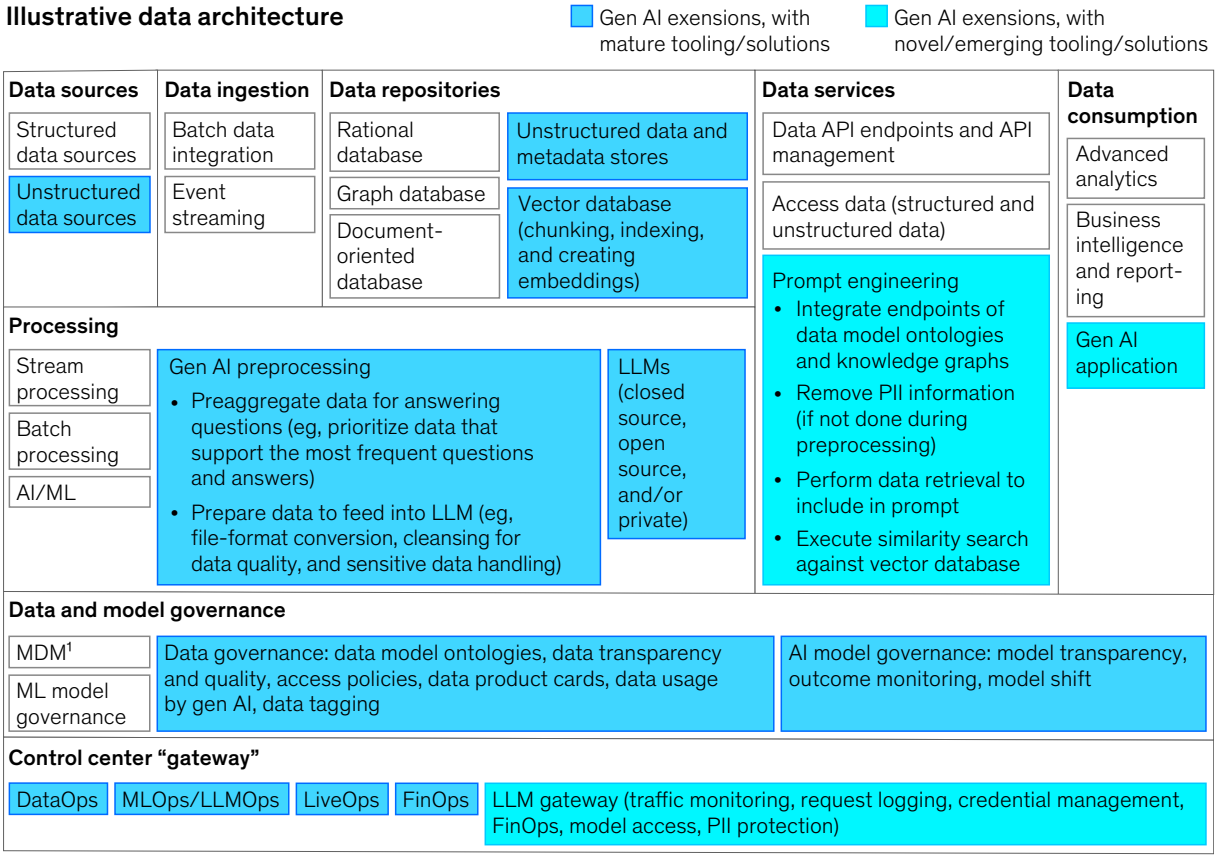
- **Data preprocessing:** Most data will need to be prepped—for example, by converting file formats and cleansing for data quality and the handling of sensitive data—so that generative AI can use the data. Preprocessed data is most often used to build prompts for generative AI models. To speed up performance, CDOs need to standardize the handling of structured and unstructured data at scale, such as ways to access underlying systems, and prioritize (or "preaggregate") the data that supports the most frequent questions and answers.

- **Vector databases:** Vectorization is a way to prioritize content and create "embeddings" (numerical representations of text meanings) in order to streamline access to context, the complementary information generative AI needs to provide accurate answers. Vector databases allow generative AI models to access just the most relevant information. Instead of providing a thousand-page PDF, for example, a vector database provides only the most relevant pages. In many cases, companies don't need to build vector databases to begin working with generative AI. They can often use existing NoSQL databases to start.

- **LLM integrations:** More-sophisticated generative AI uses require interactions with

Exhibit 2

Upgrades are needed within the existing data architecure to enable generative AI.



[1]Master data management.

McKinsey & Company

multiple systems, which creates significant challenges in connecting LLMs. Several frameworks, many of which are open source, can help facilitate these integrations (for example, LangChain or various hyperscaler offerings, such as Semantic Kernel for Azure, Bedrock for AWS, or Vertex AI for Google Cloud). CDOs will need to set guidelines for choosing which frameworks to use, define prompt templates that can be readily customized for specific purposes, and establish standardized integration patterns for how LLMs interface with source data systems.

— **Prompt engineering:** Effective prompt engineering (the process of structuring questions in a way that elicits the best response from generative AI models) relies on context. Context can be determined only from existing data and information across structured and unstructured sources. To improve output, CDOs will need to manage integration of knowledge graphs or data models and ontologies (a set of concepts in a domain that shows their properties and the relations between them) into the prompt. Since CDOs will not have ownership of many data repositories across the business, they

need to set standards and prequalify sources to ensure the data that is fed into the models follows specific protocols (for example, exposing a knowledge graph API to easily provide entities and relationships).

3. Focus on key points of the data life cycle to ensure high quality

Data quality has always been an important issue for CDOs. But the scale and scope of data that generative AI models rely on has made the "garbage in/garbage out" truism much more consequential and expensive, as training a single LLM can cost millions of dollars.[3] One reason pinpointing data quality issues is much more difficult in generative AI models than in classical ML models is because there's so much more data and much of it is unstructured, making it difficult to use existing tracking tools.

CDOs need to do two things to ensure data quality: extend their data observability programs[4] for generative AI applications to better spot quality issues, such as by setting minimum thresholds for unstructured content to be included in generative AI applications; and develop interventions across the data life cycle to fix the issues teams find, mainly in four areas:

- **Source data:** Expand the data quality framework to include measures relevant for generative AI purposes (such as bias). Ensure high-quality metadata and labels for structured and unstructured data, and regulate access to sensitive data (for example, base access on roles).

- **Preprocessing:** Ensure data is consistent and standardized and adheres to ontologies and established data models. Detect outliers and apply normalizations. Automate PII data management, and put in place guidelines for whether data should be ignored, held, redacted, quarantined, removed, masked, or synthesized.

- **Prompt:** Evaluate, measure, and track the quality of the prompt. Include high-quality metadata and lineage transparency for structured and unstructured data in the prompt.

- **Output from LLM:** Establish the necessary governance procedures to identify and resolve incorrect outputs, and use "human in the loop" to review and triage output issues. Ultimately, elevate the role of individual employees by training them to critically evaluate model outputs and be aware of the quality of input data. Supplement with an automated monitoring-and-alert capability to identify rogue behaviors.

4. Protect your sensitive data, and be ready to move quickly as regulations emerge

Some 71 percent of senior IT leaders believe generative AI technology is introducing new security risk to their data.[5] Much has been written about security and risk when it comes to generative AI, but CDOs needs to consider the data implications in three specific areas:

- **Identify and prioritize security risks to the enterprise's proprietary data.** CDOs need to assess the broad risks associated with exposing the business's data, such as the potential exposure of trade secrets when confidential and proprietary code is shared with generative AI models, and prioritize the greatest threats. Much existing data protection and cybersecurity governance can be extended to address specific generative AI risks—for example, by adding pop-up reminders whenever an engineer wants to share data with a model or by running automated scripts to ensure compliance.

- **Manage access to PII data.** CDOs need to regulate how data is detected and treated in the context of generative AI. They need to

[3] B. Urian, "NVIDIA announces $9.6 million drop in cost when using its GPUs for AI LLM training," Tech Times, May 29, 2023.
[4] Data observability programs consist of mechanisms for understanding the health and performance of the data within systems.
[5] "Top generative AI statistics for 2023," Salesforce, September 2023.

set up systems that incorporate protection tools and human interventions to ensure PII data is removed during data preprocessing and before it's used on an LLM. Using synthetic data (through data fabricators) and nonsensitive identifiers can help.

— **Track the expected surge of regulations closely.** Generative AI has acted as a catalyst to rapid movement among governments to enact new regulations, such as the European Union's AI Act, which is setting a wide array of new standards, such as having companies publish summaries of copyrighted data used for training an LLM. Data leaders must stay close to the business's risk leaders to understand new regulations and their implications for data strategy, such as the need to "untrain" models that use regulated data.

5. Build up data engineering talent

As enterprises increasingly adopt generative AI, CDOs will have to focus on the implications for talent. Some coding tasks will be done by generative AI tools—41 percent of code published on GitHub is written by AI.[6] This requires specific training on working with a generative AI "copilot"—a recent McKinsey study showed that senior engineers work more productively with a generative AI copilot than do junior engineers.[7] Data and AI academies need to incorporate generative AI training tailored to specific expertise levels.

CDOs will also need to be clear about what skills best enable generative AI. Companies need people who can integrate data sets (such as writing APIs connecting models to data sources), sequence and chain prompts, wrangle large quantities of data, apply LLMs, and work with model parameters. This means that CDOs should focus more on finding data engineers, architects, and back-end engineers, and less on hiring data scientists, whose skills will be increasingly less critical as generative AI allows people with less advanced technical capabilities to use natural language in doing basic analysis.

In the near term, talent will remain in shorter supply, and we project that the talent gap will increase further in the near future,[8] creating more incentives for CDOs to build up their training programs.

6. Use generative AI to help you manage your own data

Data leaders have a huge opportunity to harness generative AI to improve their own function. In our analysis, eight primary use cases have emerged along the entire data value chain where generative AI can both accelerate existing tasks and improve how tasks are performed (Exhibit 3).

Many vendors are already rolling out products, requiring CDOs to identify the capabilities for which they can rely on vendors and which they should build themselves. One rule of thumb is that for data governance processes that are unique to the business, it's better to build your own tool. Note that many tools and capabilities are new and may work well in experimental environments but not at scale.

7. Track rigorously and intervene quickly

There are more unknowns than knowns in the generative AI world today, and companies are still learning their way forward. It is therefore crucial for CDOs to set up systems to actively track and manage progress on their generative AI initiatives and to understand how well data is performing in supporting the business's goals.

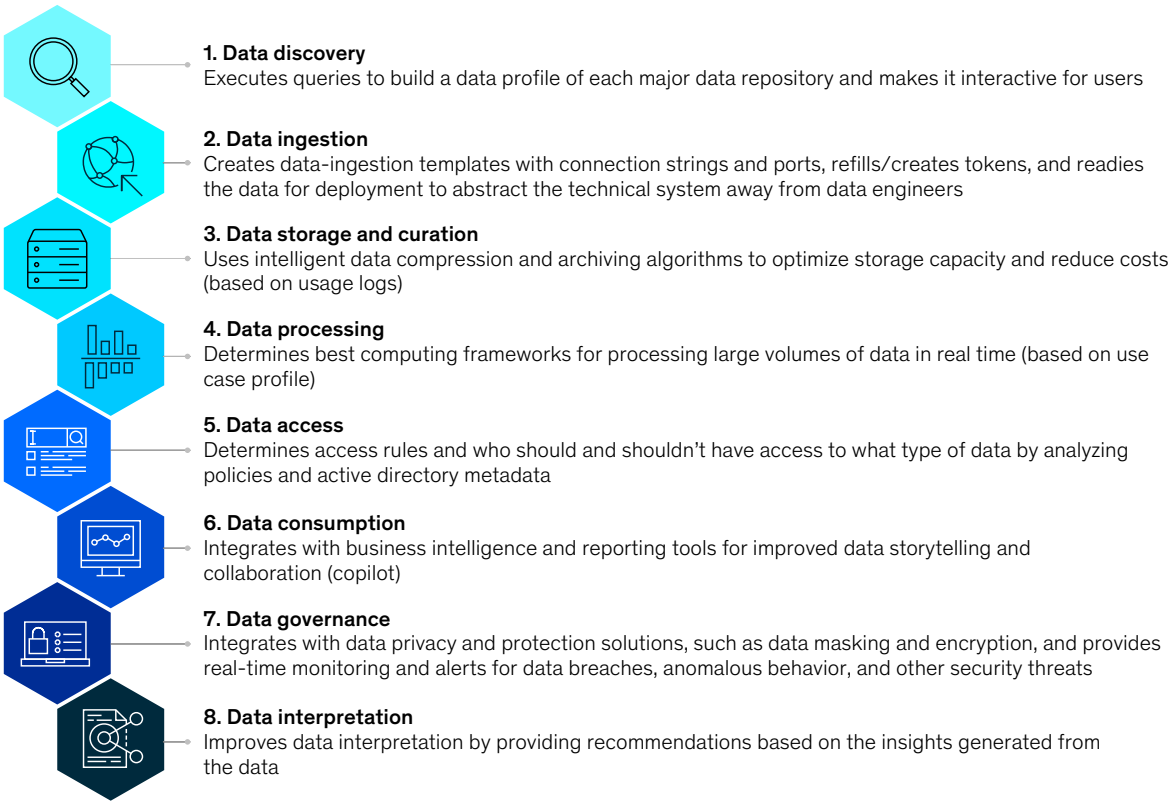
In practice, effective metrics are made up of a set of core KPIs and operational KPIs (the underlying activities that drive KPIs), which help leaders track progress and identify root causes of issues.

[6] Jose Antonio Lanz, "Stability AI CEO: There will be no (human) programmers in five years," Decrypt, June 3, 2023.
[7] "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.
[8] Michael Chui, Mena Issler, Roger Roberts, and Lareina Yee, "McKinsey Technology Trends Outlook 2023," McKinsey, July 20, 2023.

Exhibit 3

Generative AI opportunities exist to improve the entire date value chain.

Generative AI use cases along data value chain



McKinsey & Company

A core set of KPIs should include the following:

- cost of additional components, such as vector databases and consumption of LLMs as a service
- additional revenue that is enabled by the integration of specific data sources with generative AI application workflows
- time-to-market to develop a generative AI–powered application that requires access to internal data
- end-user satisfaction with how the data has improved the performance and quality of the application

Operational KPIs should include tracking which data is being used most, how models are performing, where data quality is poor, how many requests are being made against a given dataset, and which use cases are generating the most activity and value.

This information is critical in providing a fact base for leadership to not just track progress but also make rapid adjustments and trade-off decisions against other initiatives in the CDO's broader portfolio. By knowing which data sources are most used for high-value models, for example, the CDO can prioritize investments to improve data quality at those sources.

Effective investment, budgeting, and reallocation will depend on CDOs developing a FinOps-like capability to manage the entire new cost structure growing around generative AI. CDOs will need to track a new range of costs, including the number of generative AI model requests, API consumption charges from vendors (both quantity and size of calls), and compute and storage charges from cloud providers. With this information, the CDO can determine how best to optimize costs, such as routing requests by priority level or moving certain data to the cloud to cut down on networking costs.

The value of these metrics is only as great as the degree to which CDOs act on them. CDOs will need to establish data-performance metrics that can be reviewed in near real time and protocols to make rapid decisions. Effective data governance programs should remain in place but be extended to incorporate generative AI–related decisions.

Data cannot be an afterthought in generative AI. Rather, it is the core fuel that powers the ability of a business to capture value from generative AI. But businesses that want that value cannot afford CDOs who merely manage data; they need CDOs who understand how to use data to lead the business.

Joe Caserta is a partner in McKinsey's New York office, where **Kayvaun Rowshankish** is a senior partner; **Holger Harreis** is a senior partner in the Düsseldorf office, where **Asin Tavakoli** is a partner; and **Nikhil Srinidhi** is an associate partner in the Berlin office.

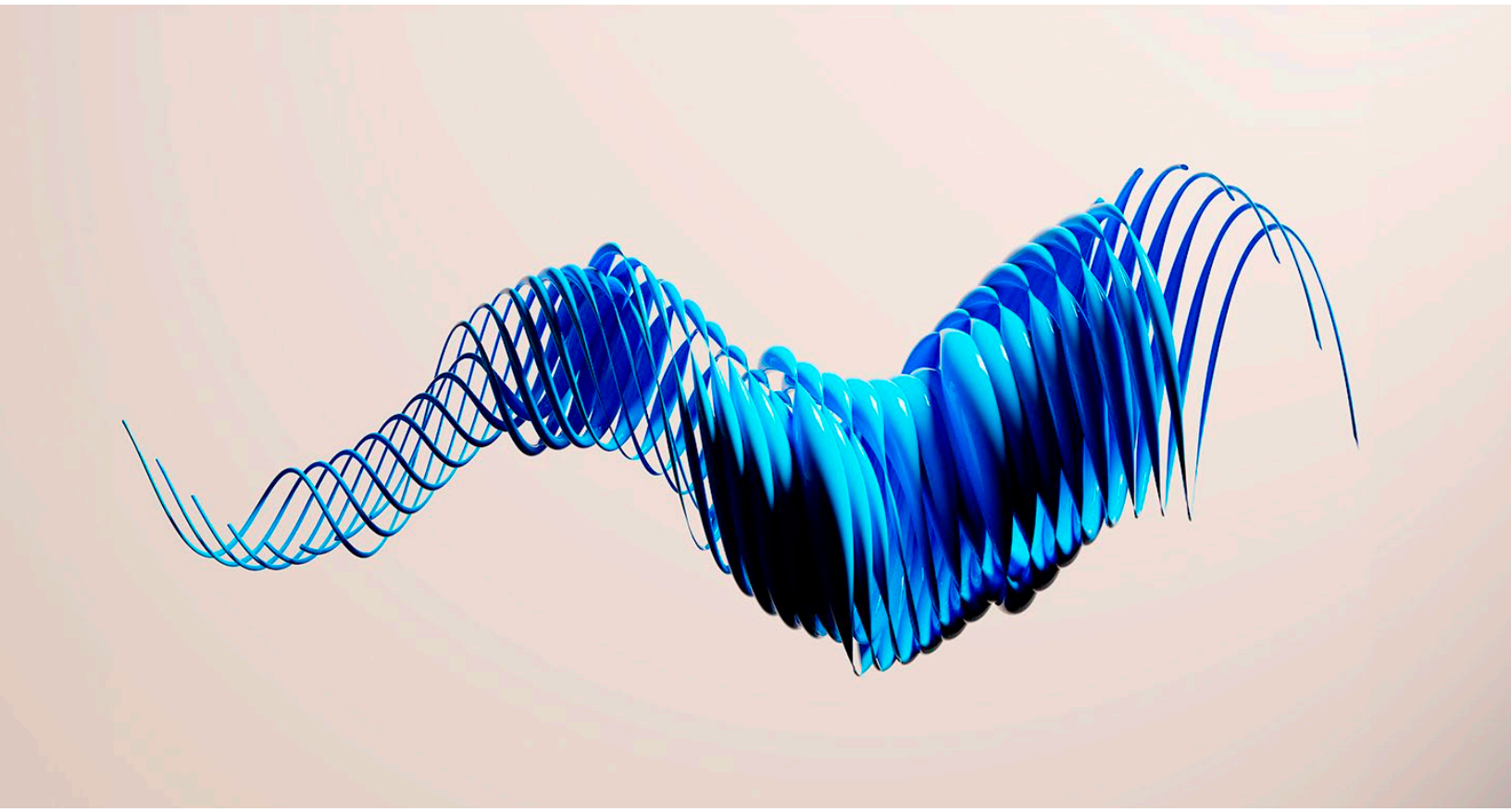
The authors wish to thank Sven Blumberg, Stephanie Brauckmann, Carlo Giovine, Jonas Heite, Vishnu Kamalnath, Simon Malberg, Rong Parnas, Bruce Philp, Adi Pradhan, Alex Singla, Saravanakumar Subramaniam, Alexander Sukharevsky, and Kevin-Morris Wigand for their contributions to this article.

McKinsey Digital

Technology's generational moment with generative AI: A CIO and CTO guide

CIOs and CTOs can take nine actions to reimagine business and technology with generative AI.

This article is a collaborative effort by Aamer Baig, Sven Blumberg, Eva Li, Douglas Merrill, Adi Pradhan, Megha Sinha, Alexander Sukharevsky, and Stephen Xu, representing views from McKinsey Digital.



July 2023

Hardly a day goes by without some new business-busting development related to generative AI surfacing in the media. The excitement is well deserved—McKinsey research estimates that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion of value annually.[1]

CIOs and chief technology officers (CTOs) have a critical role in capturing that value, but it's worth remembering we've seen this movie before. New technologies emerged—the internet, mobile, social media—that set off a melee of experiments and pilots, though significant business value often proved harder to come by. Many of the lessons learned from those developments still apply, especially when it comes to getting past the pilot stage to reach scale. For the CIO and CTO, the generative AI boom presents a unique opportunity to apply those lessons to guide the C-suite in turning the promise of generative AI into sustainable value for the business.

Through conversations with dozens of tech leaders and an analysis of generative AI initiatives at more than 50 companies (including our own), we have identified nine actions all technology leaders can take to create value, orchestrate technology and data, scale solutions, and manage risk for generative AI:

1. Move quickly to **determine the company's posture for the adoption of generative AI**, and develop practical communications to, and appropriate access for, employees.

2. Reimagine the business and **identify use cases that build value through improved productivity, growth, and new business models.** Develop a "financial AI" (FinAI) capability that can estimate the true costs and returns of generative AI.

3. **Reimagine the technology function,** and focus on quickly building generative AI capabilities in software development, accelerating technical debt reduction, and dramatically reducing manual effort in IT operations.

4. **Take advantage of existing services or adapt open-source generative AI models** to develop proprietary capabilities (building and operating your own generative AI models can cost tens to hundreds of millions of dollars, at least in the near term).

5. **Upgrade your enterprise technology architecture to integrate and manage generative AI models** and orchestrate how they operate with each other and existing AI and machine learning (ML) models, applications, and data sources.

6. **Develop a data architecture to enable access to quality data** by processing both structured and unstructured data sources.

7. **Create a centralized, cross-functional generative AI platform team** to provide approved models to product and application teams on demand.

8. Invest in upskilling key roles—software developers, data engineers, MLOps engineers, and security experts—as well as the broader nontech workforce. But you need to **tailor the training programs by roles and proficiency levels** due to the varying impact of generative AI.

9. **Evaluate the new risk landscape and establish ongoing mitigation practices** to address models, data, and policies.

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

1. Determine the company's posture for the adoption of generative AI

As use of generative AI becomes increasingly widespread, we have seen CIOs and CTOs respond by blocking employee access to publicly available applications to limit risk. In doing so, these companies risk missing out on opportunities for innovation, with some employees even perceiving these moves as limiting their ability to build important new skills.

Instead, CIOs and CTOs should work with risk leaders to balance the real need for risk mitigation with the importance of building generative AI skills in the business. This requires establishing the company's posture regarding generative AI by building consensus around the levels of risk with which the business is comfortable and how generative AI fits into the business's overall strategy. This step allows the business to quickly determine company-wide policies and guidelines.

Once policies are clearly defined, leaders should communicate them to the business, with the CIO and CTO providing the organization with appropriate access and user-friendly guidelines. Some companies have rolled out firmwide communications about generative AI, provided broad access to generative AI for specific user groups, created pop-ups that warn users any time they input internal data into a model, and built a guidelines page that appears each time users access a publicly available generative AI service.

2. Identify use cases that build value through improved productivity, growth, and new business models

CIOs and CTOs should be the antidote to the "death by use case" frenzy that we already see in many companies. They can be most helpful by working with the CEO, CFO, and other business leaders to think through how generative AI challenges existing business models, opens doors to new ones, and creates new sources of value. With a deep understanding of the technical possibilities, the CIO and CTO should identify the most valuable opportunities and issues across the company that can benefit from generative AI—and those that can't. In some cases, generative AI is *not* the best option.

McKinsey research, for example, shows generative AI can lift productivity for certain marketing use cases (for example, by analyzing unstructured and abstract data for customer preference) by roughly 10 percent and customer support (for example, through intelligent bots) by up to 40 percent.[2] The CIO and CTO can be particularly helpful in developing a perspective on how best to cluster use cases either by domain (such as customer journey or business process) or use case type (such as creative content creation or virtual agents) so that generative AI will have the most value. Identifying opportunities won't be the most strategic task—there are many generative AI use cases out there—but, given initial limitations of talent and capabilities, the CIO and CTO will need to provide feasibility and resource estimates to help the business sequence generative AI priorities.

Providing this level of counsel requires tech leaders to work with the business to develop a FinAI capability to estimate the true costs and returns on generative AI initiatives. Cost calculations can be particularly complex because the unit economics must account for multiple model and vendor costs, model interactions (where a query might require input from multiple models, each with its own fee), ongoing usage fees, and human oversight costs.

3. Reimagine the technology function

Generative AI has the potential to completely remake how the tech function works. CIOs and CTOs need to make a comprehensive review of the potential impact of generative AI on all areas of tech,

but it's important to take action quickly to build experience and expertise. There are three areas where they can focus their initial energies:

— **Software development:** McKinsey research shows generative AI coding support can help software engineers develop code 35 to 45 percent faster, refactor code 20 to 30 percent faster, and perform code documentation 45 to 50 percent faster.[3] Generative AI can also automate the testing process and simulate edge cases, allowing teams to develop more-resilient software prior to release, and accelerate the onboarding of new developers (for example, by asking generative AI questions about a code base). Capturing these benefits will require extensive training (see more in action 8) and automation of integration and deployment pipelines through DevSecOps practices to manage the surge in code volume.

— **Technical debt:** Technical debt can account for 20 to 40 percent of technology budgets and significantly slow the pace of development.[4] CIOs and CTOs should review their tech-debt balance sheets to determine how generative AI capabilities such as code refactoring, code translation, and automated test-case generation can accelerate the reduction of technical debt.

— **IT operations (ITOps):** CIOs and CTOs will need to review their ITOps productivity efforts to determine how generative AI can accelerate processes. Generative AI's capabilities are particularly helpful in automating such tasks as password resets, status requests, or basic diagnostics through self-serve agents; accelerating triage and resolution through improved routing; surfacing useful context, such as topic or priority, and generating suggested responses; improving observability through analysis of vast streams of logs to identify events that truly require attention; and developing documentation, such as standard operating procedures, incident postmortems, or performance reports.

4. Take advantage of existing services or adapt open-source generative AI models

A variation of the classic "rent, buy, or build" decision exists when it comes to strategies for developing generative AI capabilities. The basic rule holds true: a company should invest in a generative AI capability where it can create a proprietary advantage for the business and access existing services for those that are more like commodities.

The CIO and CTO can think through the implications of these options as three archetypes:

— **Taker**—uses publicly available models through a chat interface or an API, with little or no customization. Good examples include off-the-shelf solutions to generate code (such as GitHub Copilot) or to assist designers with image generation and editing (such as Adobe Firefly). This is the simplest archetype in terms of both engineering and infrastructure needs and is generally the fastest to get up and running. These models are essentially commodities that rely on feeding data in the form of prompts to the public model.

— **Shaper**—integrates models with internal data and systems to generate more customized results. One example is a model that supports sales deals by connecting generative AI tools to customer relationship management (CRM) and financial systems to incorporate customers' prior sales and engagement history. Another is fine-tuning the model with internal company documents and chat history to act as an assistant to a customer support agent. For companies that are looking to

[2] Ibid.
[3] Begum Karaci Deniz, Martin Harrysson, Alharith Hussin, and Shivam Srivastava, "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.
[4] Vishal Dalal, Krish Krishnakanthan, Björn Münstermann, and Rob Patenge, "Tech debt: Reclaiming tech equity," McKinsey, October 6, 2020.

scale generative AI capabilities, develop more proprietary capabilities, or meet higher security or compliance needs, the Shaper archetype is appropriate.

There are two common approaches for integrating data with generative AI models in this archetype. One is to "bring the model to the data," where the model is hosted on the organization's infrastructure, either on-premises or in the cloud environment. Cohere, for example, deploys foundation models on clients' cloud infrastructure, reducing the need for data transfers. The other approach is to "bring data to the model," where an organization can aggregate its data and deploy a copy of the large model on cloud infrastructure. Both approaches achieve the goal of providing access to the foundation models, and choosing between them will come down to the organization's workload footprint.

— **Maker**—builds a foundation model to address a discrete business case. Building a foundation model is expensive and complex, requiring huge volumes of data, deep expertise, and massive compute power. This option requires a substantial one-off investment—tens or even hundreds of millions of dollars—to build the model and train it. The cost depends on various factors, such as training infrastructure, model architecture choice, number of model parameters, data size, and expert resources.

Each archetype has its own costs that tech leaders will need to consider (Exhibit 1). While new developments, such as efficient model training approaches and lower graphics processing unit (GPU) compute costs over time, are driving costs down, the inherent complexity of the Maker archetype means that few organizations will adopt it in the short term. Instead, most will turn to some combination of Taker, to quickly access a commodity service, and Shaper, to build a proprietary capability on top of foundation models.

5. Upgrade your enterprise technology architecture to integrate and manage generative AI models

Organizations will use many generative AI models of varying size, complexity, and capability. To generate value, these models need to be able to work both together and with the business's existing systems or applications. For this reason, building a separate tech stack for generative AI creates more complexities than it solves. As an example, we can look at a consumer querying customer service at a travel company to resolve a booking issue (Exhibit 2). In interacting with the customer, the generative AI model needs to access multiple applications and data sources.

For the Taker archetype, this level of coordination isn't necessary. But for companies looking to scale the advantages of generative AI as Shapers or Makers, CIOs and CTOs need to upgrade their technology architecture. The prime goal is to integrate generative AI models into internal systems and enterprise applications and to build pipelines to various data sources. Ultimately, it's the maturity of the business's enterprise technology architecture that allows it to integrate and scale its generative AI capabilities.

Recent advances in integration and orchestration frameworks, such as LangChain and LlamaIndex, have significantly reduced the effort required to connect different generative AI models with other applications and data sources. Several integration patterns are also emerging, including those that enable models to call APIs when responding to a user query—GPT-4, for example, can invoke functions—and provide contextual data from an external dataset as part of a user query, a technique known as retrieval augmented generation. Tech leaders will need to define reference architectures and standard integration patterns for their organization (such as standard API formats and parameters that identify the user and the model invoking the API).

Exhibit 1

Each archetype has its own costs.

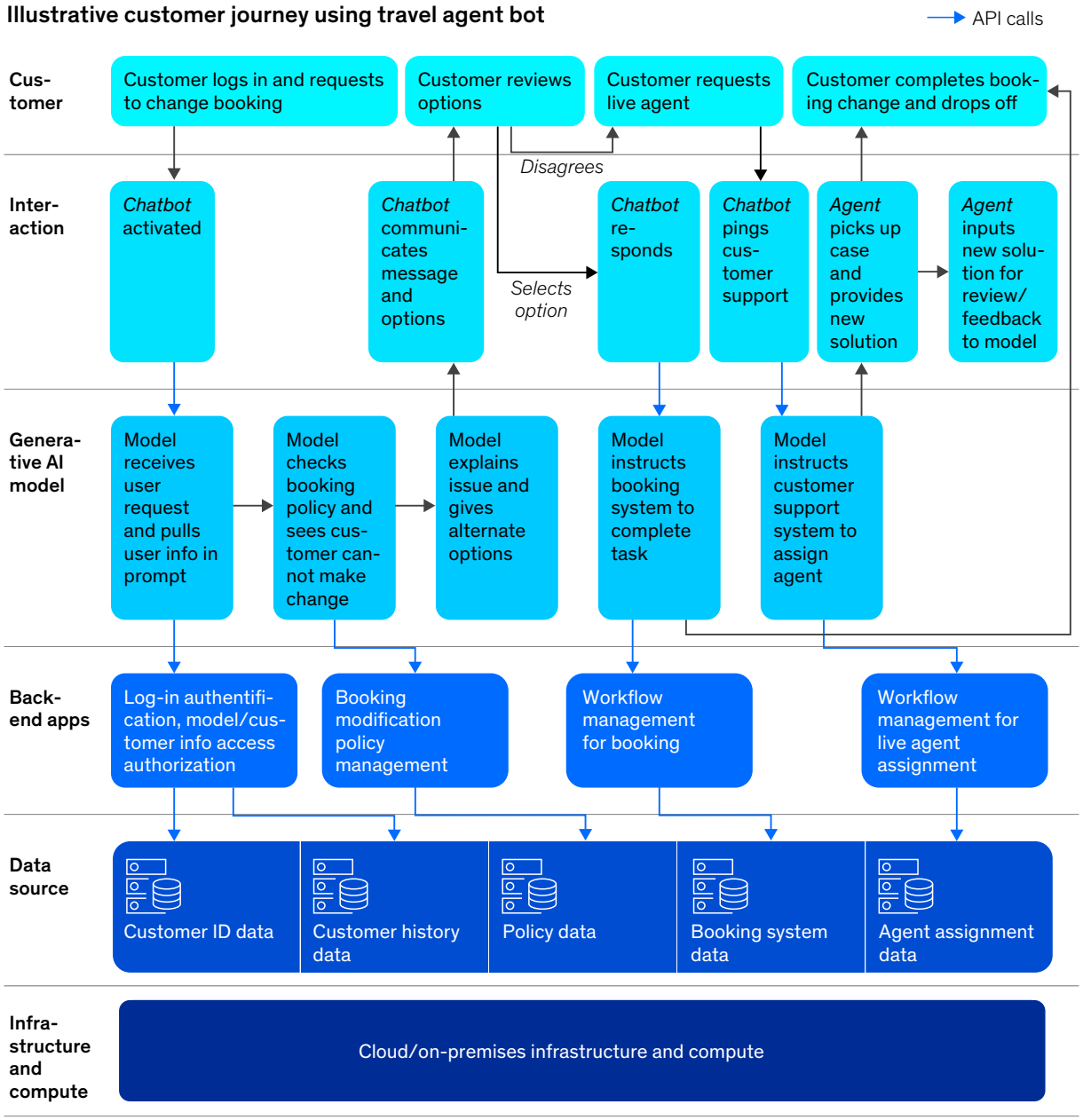
Archetype	Example use cases	Estimated total cost of ownership
Taker	— Off-the-shelf coding assistant for software developers — General-purpose customer service chatbot with prompt engineering only and text chat only	**~ $0.5 million to $2.0 million, onetime** — Off-the-shelf coding assistant: ~$0.5 million for integration. Costs include a team of 6 working for 3 to 4 months. — General-purpose customer service chatbot: ~$2.0 million for building plug-in layer on top of third-party model API. Costs include a team of 8 working for 9 months. **~ $0.5 million, recurring annually** — Model inference: • Off-the-shelf coding assistant: ~$0.2 million annually per 1,000 daily users • General-purpose customer service chatbot: ~$0.2 million annually, assuming 1,000 customer chats per day and 10,000 tokens per chat — Plug-in-layer maintenance: up to ~$0.2 million annually, assuming 10% of development cost.
Shaper	— Customer service chatbot fine-tuned with sector-specific knowledge and chat history	**~ $2.0 million to $10.0 million, onetime unless model is fine-tuned further** — Data and model pipeline building: ~$0.5 million. Costs include 5 to 6 machine learning engineers and data engineers working for 16 to 20 weeks to collect and label data and perform data ETL.[1] — Model fine-tuning[2]: ~$0.1 million to $6.0 million per training run[3] • Lower end: costs include compute and 2 data scientists working for 2 months • Upper end: compute based on public closed-source model fine-tuning cost — Plug-in-layer building: ~$1.0 million to $3.0 million. Costs include a team of 6 to 8 working for 6 to 12 months. **~ 0.5 million to $1.0 million, recurring annually** — Model inference: up to ~$0.5 million recurring annually. Assume 1,000 chats daily with both audio and texts. — Model maintenance: ~$0.5 million. Assume $100,000 to $250,000 annually for MLOps platform[4] and 1 machine learning engineer spending 50% to 100% of their time monitoring model performance. — Plug-in-layer maintenance: up to ~$0.3 million recurring annually, assuming 10% of development cost.
Maker	— Foundation model trained for assisting in patient diagnosis	**~ $5.0 million to $200.0 million, onetime unless model is fine-tuned or retrained** — Model development: ~$0.5 million. Costs include 4 data scientists spending 3 to 4 months on model design, development, and evaluation leveraging existing research. — Data and model pipeline: ~$0.5 million to $1.0 million. Costs include 6 to 8 machine learning engineers and data engineers working for ~12 weeks to collect data and perform data ETL.[1] — Model training[5]: ~$4.0 million to $200.0 million per training run.[3] Costs include compute and labor cost of 4 to 6 data scientists working for 3 to 6 months. — Plug-in-layer building: ~$1.0 million to $3.0 million. Costs include a team of 6 to 8 working 6 to 12 months. **~ $1.0 million to $5.0 million, recurring annually** — Model inference: ~$0.1 million to $1.0 million annually per 1,000 users. Assume each physician sees 20 to 25 patients per day and patient speaks for 6 to 25 minutes per visit. — Model maintenance: ~$1.0 million to $4.0 million recurring annually. Assume $250,000 annually for MLOps platform[4] and 3 to 5 machine learning engineers to monitor model performance. — Plug-in-layer maintenance: up to ~$0.3 million recurring annually, assuming 10% of development cost.

Note: Through engineering optimizations, the economics of generative AI are evolving rapidly, and these are high-level estimates based on total cost of ownership (resources, model training, etc) as of mid-2023.

[1] Extract, transform, and load.
[2] Model is fine-tuned on dataset consisting of ~100,000 pages of sector-specific documents and 5 years of chat history from ~1,000 customer representatives, which is ~48 billion tokens. Lower end cost consists of 1% parameters retrained on open-source models (eg, LLaMA) and upper end on closed-source models. Chatbot can be accessed via both text and audio.
[3] Model is optimized after each training run based on use of hyperparameters, dataset, and model architecture. Model may be refreshed periodically when needed (eg, with fresh data).
[4] Gilad Shaham, "Build or buy your MLOps platform: Main considerations," LinkedIn, November 3, 2021.
[5] Model is trained on 65 billion to 1 trillion parameters and dataset of 1.2 to 2.4 trillion tokens. The tool can be accessed via both text and audio.

Exhibit 2

Generative AI is integrated at key touchpoints to enable a tailored customer journey.



There are five key elements that need to be incorporated into the technology architecture to integrate generative AI effectively (Exhibit 3):

— **Context management and caching** to provide models with relevant information from enterprise data sources. Access to relevant data at the right time is what allows the model to understand the context and produce compelling outputs. Caching stores results to frequently asked questions to enable faster and cheaper responses.

— **Policy management** to ensure appropriate access to enterprise data assets. This control ensures that HR's generative AI models that include employee compensation details, for example, cannot be accessed by the rest of the organization.

— **Model hub,** which contains trained and approved models that can be provisioned on demand and acts as a repository for model checkpoints, weights, and parameters.

— **Prompt library,** which contains optimized instructions for the generative AI models, including prompt versioning as models are updated.

— **MLOps platform,** including upgraded MLOps capabilities, to account for the complexity of generative AI models. MLOps pipelines, for example, will need to include instrumentation to measure task-specific performance, such as measuring a model's ability to retrieve the right knowledge.

In evolving the architecture, CIOs and CTOs will need to navigate a rapidly growing ecosystem of generative AI providers and tooling. Cloud providers provide extensive access to at-scale hardware and foundation models, as well as a proliferating set of services. MLOps and model hub providers, meanwhile, offer the tools, technologies, and practices to adapt a foundation model and deploy it into production, while other companies provide applications directly accessed by users built on top of foundation models to perform specific tasks. CIOs and CTOs will need to assess how these various capabilities are assembled and integrated to deploy and operate generative AI models.

6. Develop a data architecture to enable access to quality data

The ability of a business to generate and scale value, including cost reductions and improved data and knowledge protections, from generative AI models will depend on how well it takes advantage of its own data. Creating that advantage relies on a data architecture that connects generative AI models to internal data sources, which provide context or help fine-tune the models to create more relevant outputs.

In this context, CIOs, CTOs, and chief data officers need to work closely together to do the following:

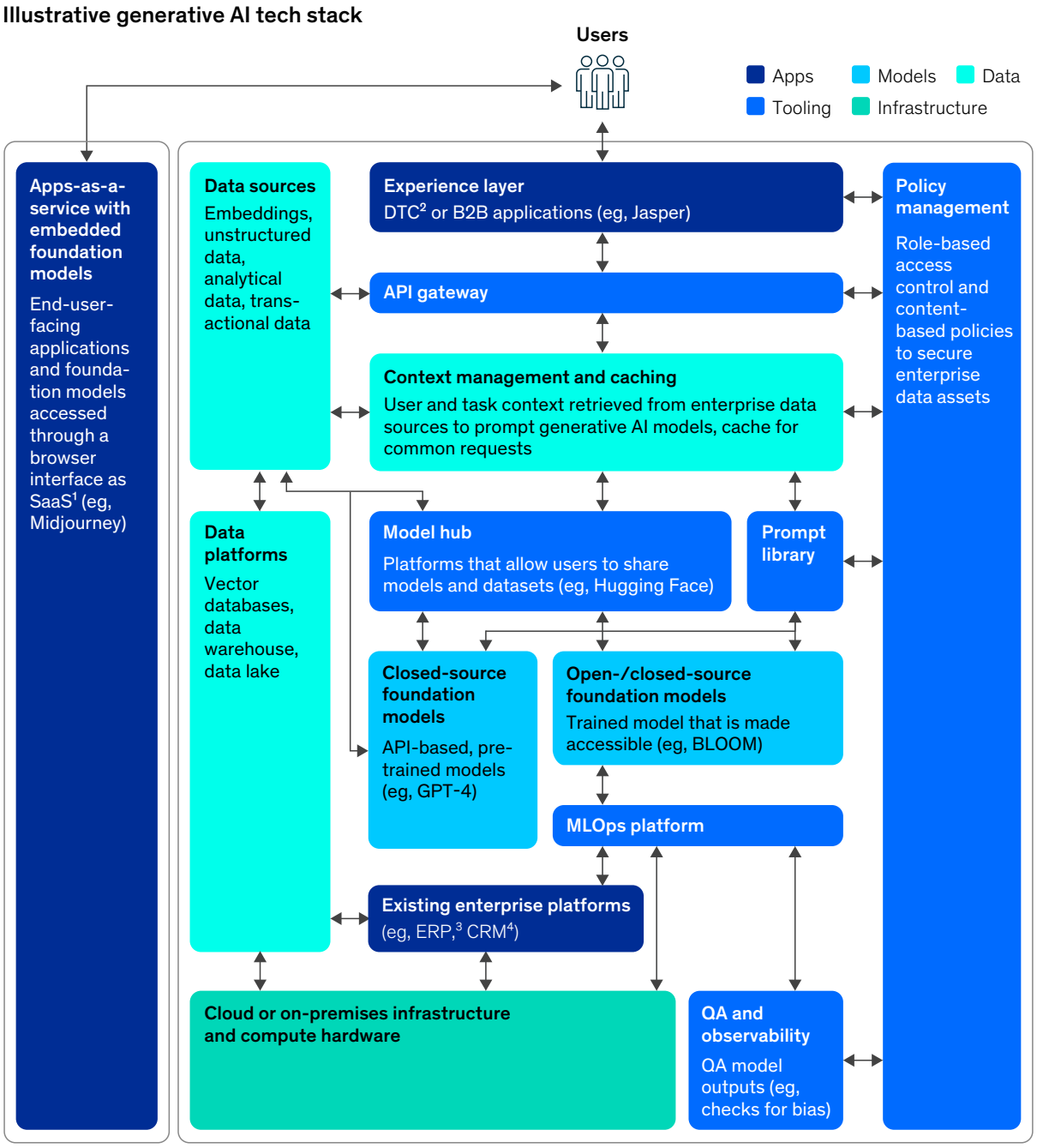
— Categorize and organize data so it can be used by generative AI models. Tech leaders will need to develop a comprehensive data architecture that encompasses both structured and unstructured data sources. This requires putting in place standards and guidelines to optimize data for generative AI use—for example, by augmenting training data with synthetic samples to improve diversity and size; converting media types into standardized data formats; adding metadata to improve traceability and data quality; and updating data.

— Ensure existing infrastructure or cloud services can support the storage and handling of the vast volumes of data needed for generative AI applications.

— Prioritize the development of data pipelines to connect generative AI models to relevant data sources that provide "contextual understanding." Emerging approaches include the use of vector databases to store and retrieve embeddings (specially formatted

Exhibit 3

The tech stack for generative AI is emerging.



[1]Software as a service.
[2]Direct to consumer.
[3]Enterprise resource planning.
[4]Customer relationship management.

knowledge) as input for generative AI models as well as in-context learning approaches, such as "few shot prompting," where models are provided with examples of good answers.

7. Create a centralized, cross-functional generative AI platform team

Most tech organizations are on a journey to a product and platform operating model. CIOs and CTOs need to integrate generative AI capabilities into this operating model to build on the existing infrastructure and help to rapidly scale adoption of generative AI. The first step is setting up a generative AI platform team whose core focus is developing and maintaining a platform service where approved generative AI models can be provisioned on demand for use by product and application teams. The platform team also defines protocols for how generative AI models integrate with internal systems, enterprise applications, and tools, and also develops and implements standardized approaches to manage risk, such as responsible AI frameworks.

CIOs and CTOs need to ensure that the platform team is staffed with people who have the right skills. This team requires a senior technical leader who acts as the general manager. Key roles include software engineers to integrate generative AI models into existing systems, applications, and tools; data engineers to build pipelines that connect models to various systems of record and data sources; data scientists to select models and engineer prompts; MLOps engineers to manage deployment and monitoring of multiple models and model versions; ML engineers to fine-tune models with new data sources; and risk experts to manage security issues such as data leakage, access controls, output accuracy, and bias. The exact composition of the platform team will depend on the use cases being served across the enterprise. In some instances, such as creating a customer-facing chatbot, strong product management and user experience (UX) resources will be required.

Realistically, the platform team will need to work initially on a narrow set of priority use cases, gradually expanding the scope of their work as they build reusable capabilities and learn what works best. Technology leaders should work closely with business leads to evaluate which business cases to fund and support.

8. Tailor upskilling programs by roles and proficiency levels

Generative AI has the potential to massively lift employees' productivity and augment their capabilities. But the benefits are unevenly distributed depending on roles and skill levels, requiring leaders to rethink how to build the actual skills people need.

Our latest empirical research using the generative AI tool GitHub Copilot, for example, helped software engineers write code 35 to 45 percent faster.[5] The benefits, however, varied. Highly skilled developers saw gains of up to 50 to 80 percent, while junior developers experienced a 7 to 10 percent ***decline*** in speed. That's because the output of the generative AI tools requires engineers to critique, validate, and improve the code, which inexperienced software engineers struggle to do. Conversely, in less technical roles, such as customer service, generative AI helps low-skill workers significantly, with productivity increasing by 14 percent and staff turnover dropping as well, according to one study.[6]

These disparities underscore the need for technology leaders, working with the chief human resources officer (CHRO), to rethink their talent management strategy to build the workforce of the future. Hiring a core set of top generative AI talent will be important, and, given the increasing scarcity and strategic importance of that talent, tech leaders should put in place retention mechanisms, such as competitive salaries and opportunities to be involved in important strategic work for the business.

Tech leaders, however, cannot stop at hiring. Because nearly every existing role will be affected

[5] "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.

by generative AI, a crucial focus should be on upskilling people based on a clear view of what skills are needed by role, proficiency level, and business goals. Let's look at software developers as an example. Training for novices needs to emphasize accelerating their path to become top code reviewers in addition to code generators. Similar to the difference between writing and editing, code review requires a different skill set. Software engineers will need to understand what good code looks like; review the code created by generative AI for functionality, complexity, quality, and readability; and scan for vulnerabilities while ensuring they do not themselves introduce quality or security issues in the code. Furthermore, software developers will need to learn to *think* differently when it comes to coding, by better understanding user intent so they can create prompts and define contextual data that help generative AI tools provide better answers.

Beyond training up tech talent, the CIO and CTO can play an important role in building generative AI skills among nontech talent as well. Besides understanding how to use generative AI tools for such basic tasks as email generation and task management, people across the business will need to become comfortable using an array of capabilities to improve performance and outputs. The CIO and CTO can help adapt academy models to provide this training and corresponding certifications.

The decreasing value of inexperienced engineers should accelerate the move away from a classic talent pyramid, where the greatest number of people are at a junior level, to a structure more like a diamond, where the bulk of the technical workforce is made up of experienced people. Practically speaking, that will mean building the skills of junior employees as quickly as possible while reducing roles dedicated to low-complexity manual tasks (such as writing unit tests).

9. Evaluate the new risk landscape and establish ongoing mitigation practices

Generative AI presents a fresh set of ethical questions and risks, including "hallucinations," whereby the generative AI model presents an incorrect response based on the highest-probability response; the accidental release of confidential personally identifiable information; inherent bias in the large datasets the models use; and high degrees of uncertainty related to intellectual property (IP). CIOs and CTOs will need to become fluent in ethics, humanitarian, and compliance issues to adhere not just to the letter of the law (which will vary by country) but also to the spirit of responsibly managing their business's reputation.

Addressing this new landscape requires a significant review of cyber practices and updating the software development process to evaluate risk and identify mitigation actions before model development begins, which will both reduce issues and ensure the process doesn't slow down. Proven risk-mitigation actions for hallucinations can include adjusting the level of creativity (known as the "temperature") of a model when it generates responses; augmenting the model with relevant internal data to provide more context; using libraries that impose guardrails on what can be generated; using "moderation" models to check outputs; and adding clear disclaimers. Early generative AI use cases should focus on areas where the cost of error is low, to allow the organization to work through inevitable setbacks and incorporate learnings.

To protect data privacy, it will be critical to establish and enforce sensitive data tagging protocols, set up data access controls in different domains (such as HR compensation data), add extra protection when data is used externally,

[6] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, *Generative AI at work*, National Bureau of Economic Research (NBER) working paper, number 31161, April 2023.

and include privacy safeguards. For example, to mitigate access control risk, some organizations have set up a policy-management layer that restricts access by role once a prompt is given to the model. To mitigate risk to intellectual property, CIOs and CTOs should insist that providers of foundation models maintain transparency regarding the IP (data sources, licensing, and ownership rights) of the datasets used.

Generative AI is poised to be one of the fastest-growing technology categories we've ever seen. Tech leaders cannot afford unnecessary delays in defining and shaping a generative AI strategy. While the space will continue to evolve rapidly, these nine actions can help CIOs and CTOs responsibly and effectively harness the power of generative AI at scale.

Aamer Baig is a senior partner in McKinsey's Chicago office; **Sven Blumberg** is a senior partner in the Düsseldorf office; **Eva Li** is a consultant in the Bay Area office, where **Megha Sinha** is a partner; **Douglas Merrill** is a partner in the Southern California office; **Adi Pradhan** and **Stephen Xu** are associate partners in the Toronto office; and **Alexander Sukharevsky** is a senior partner in the London office.

The authors wish to thank Stephanie Brauckmann, Anusha Dhasarathy, Martin Harrysson, Klemens Hjartar, Alharith Hussin, Naufal Khan, Sam Nie, Chandrasekhar Panda, Henning Soller, Nikhil Srinidhi, Asin Tavakoli, Niels Van der Wildt, and Anna Wiesinger for their contributions to this article.

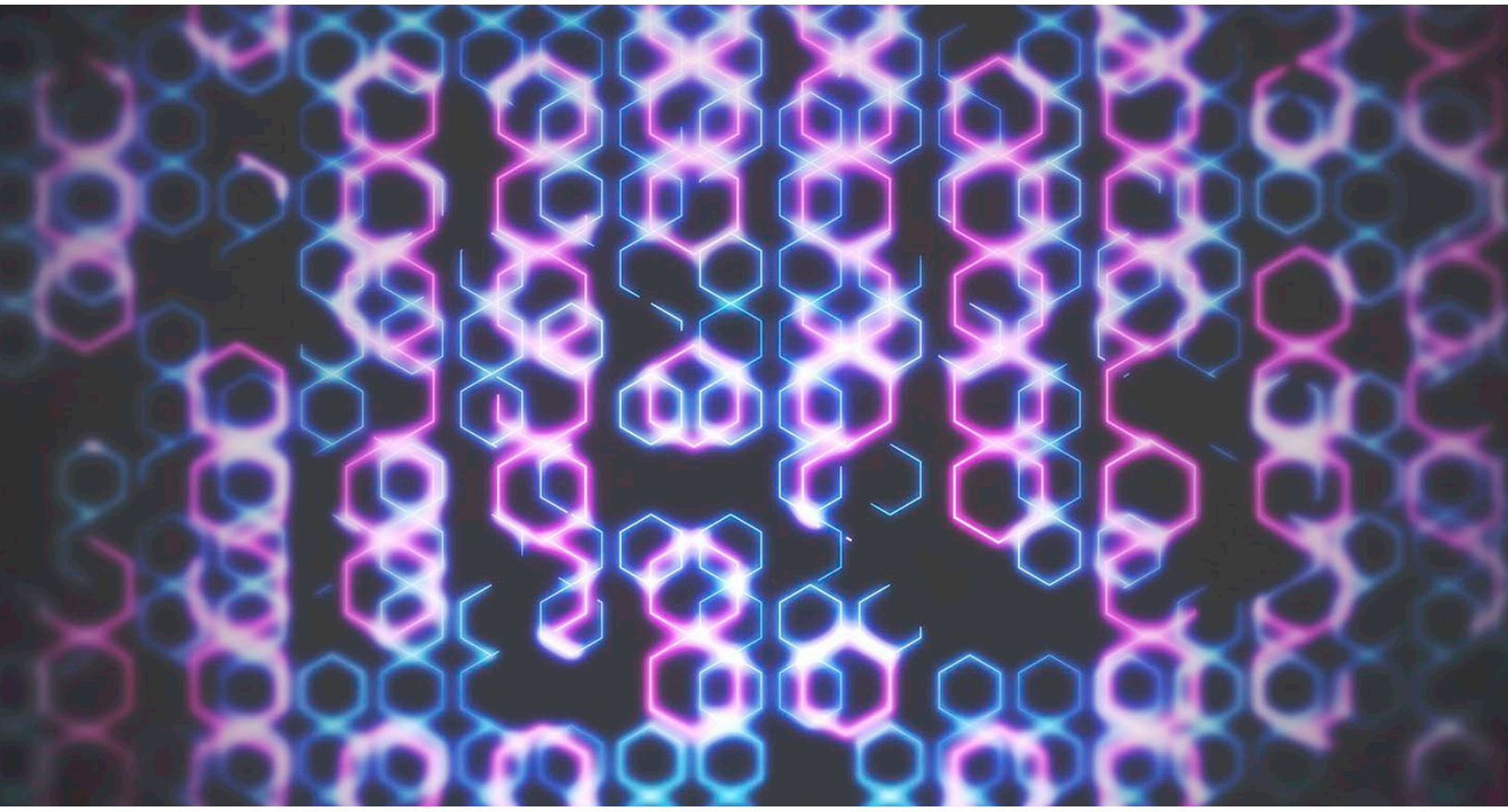
McKinsey & Company

Risk & Resilience Practice

As gen AI advances, regulators—and risk functions—rush to keep pace

AI and its supercharged breakthrough, generative AI, are all about rapid advancements, and rule makers are under pressure to keep up.

This article is a collaborative effort by Andreas Kremer, Angela Luget, Daniel Mikkelsen, Henning Soller, Malin Strandell-Jansson, and Sheila Zingg, representing views from McKinsey's Risk & Resilience Practice.



December 2023

The rapid advancement of generative AI (gen AI) has regulators around the world racing to understand, control, and guarantee the safety of the technology—all while preserving its potential benefits. Across industries, gen AI adoption has presented a new challenge for risk and compliance functions: how to balance use of this new technology amid an evolving—and uneven—regulatory framework.

As governments and regulators try to define what such a control environment should look like, the developing approaches are fragmented and often misaligned, making it difficult for organizations to navigate and causing substantial uncertainty.

In this article, we explain the risks of AI and gen AI and why the technology has drawn regulatory scrutiny. We also offer a strategic road map to help risk functions navigate the uneven and changing rule-making landscape—which is focused not only on gen AI but all artificial intelligence.

Why does gen AI need regulation?

AI's breakthrough advancement, gen AI, has quickly captured the interest of the public, with ChatGPT becoming one of the fastest-growing platforms ever, reaching one million users in just five days. The acceleration comes as no surprise given the wide range of gen AI use cases, which promise increased productivity, expedited access to knowledge, and an expected total economic impact of $2.6 trillion to $4.4 trillion annually.[1]

There is, however, an economic incentive to getting AI and gen AI adoption right. Companies developing these systems may face consequences if the platforms they develop are not sufficiently polished. And a misstep can be costly. Major gen AI companies, for example, have lost significant market value when their platforms were found hallucinating (when AI generates false or illogical information).

The proliferation of gen AI has increased the visibility of risks. Key gen AI concerns include how the technology's models and systems are developed and how the technology is used.

Generally, there are concerns about a potential lack of transparency in the functioning of gen AI systems, the data used to train them, issues of bias and fairness, potential intellectual property infringements, possible privacy violations, third-party risk, as well as security concerns.

Add disinformation to these concerns, such as erroneous or manipulated output and harmful or malicious content, and it is no wonder regulators are seeking to mitigate potential harms. Regulators seek to establish legal certainty for companies engaged in the development or use of gen AI. Meanwhile, rule makers want to encourage innovation without fear of unknown repercussions.

The goal is to establish harmonized international regulatory standards that would stimulate international trade and data transfers. In pursuit of this goal, a consensus has been reached: the gen AI development community has been at the forefront of advocating for some regulatory control over the technology's development as soon as possible. The question at hand is not whether to proceed with regulations, but rather how.

The current international regulatory landscape for AI

While no country has passed comprehensive AI or gen AI regulation to date, leading legislative efforts include those in Brazil, China, the European Union, Singapore, South Korea, and the United States. The approaches taken by the different countries vary from broad AI regulation supported by existing data protection and cybersecurity regulations (the European Union and South Korea) to sector-specific laws (the United States) and more general principles or guidelines-based approaches (Brazil, Singapore, and the United States). Each approach has its own benefits and drawbacks, and some markets will move from principles-based guidelines to strict legislation over time (Exhibit 1).

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

Exhibit 1

Regulations related to AI governance vary around the world.

As of November 2023, nonexhaustive

Source: OECD; McKinsey analysis

McKinsey & Company

While the approaches vary, common themes in the regulatory landscape have emerged globally:

— *Transparency.* Regulators are seeking traceability and clarity of AI output. Their goal is to ensure that users are informed when they engage with any AI system and to provide them with information about their rights and about the capabilities and limitations of the system.

— *Human agency and oversight.* Ideally, AI systems should be developed and used as tools that serve people, uphold human dignity and personal autonomy, and function in a way that can be appropriately controlled and overseen by humans.

— *Accountability.* Regulators want to see mechanisms that ensure awareness of responsibilities, accountability, and potential redress regarding AI systems. In practice, they are seeking top management buy-in, organization-wide education, and awareness of individual responsibility.

— *Technical robustness and safety.* Rule makers are seeking to minimize unintended and unexpected harm by ensuring that AI systems are robust, meaning they operate as expected, remain stable, and can rectify user errors. They should have fallback solutions and remediation to address any failures to meet these criteria, and they should be resilient against attempts to manipulate the system by malicious third parties.

— *Diversity, nondiscrimination, and fairness.* Another goal for regulators is to ensure that AI systems are free of bias and that the output does not result in discrimination or unfair treatment of people.

— *Privacy and data governance.* Regulators want to see development and usage of AI systems that follow existing privacy and data protection rules while processing data that meets high standards in quality and integrity.

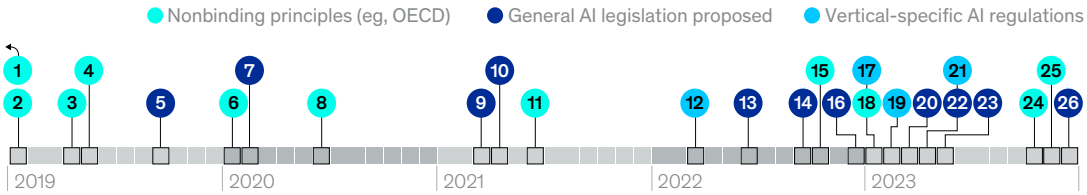
— *Social and environmental well-being.* There is a strong desire to ensure that all AI is sustainable, environmentally friendly (for instance, in its energy use), and beneficial to all people, with ongoing monitoring and assessing of the long-term effects on individuals, society, and democracy.

Despite some commonality in the guiding principles of AI, the implementation and exact wording vary by regulator and region. Many rules are still new and, thus, prone to frequent updates (Exhibit 2). This makes it challenging for organizations to navigate regulations while planning long-term AI strategies.

Exhibit 2

AI governance–related policy and regulatory efforts are under way globally.

Examples by type of policy or effort and when proposed; nonexhaustive



2019 and earlier

- **1. Sept 2017** South Korea Ethical Guidelines for Intelligent Information Technology
- **2. Jan 2019** Singapore Model AI Governance Framework, first edition
- **3. Apr 2019** EU Ethics Guidelines for Trustworthy AI
- **4. May 2019** OECD AI Principles
- **5. Sept 2019** Bill establishing the principles for the use of AI in Brazil

2020

- **6. Jan 2020** Singapore Model AI Governance Framework, second edition
- **7. Feb 2020** Bill establishing the fundamental principles and guidelines for the development and application of AI in Brazil
- **8. June 2020** South Korea Framework Act on Intelligent Informatization

2021

- **9. Mar 2021** Bill providing for the ethical framework and guidelines that underlie the development and use of AI in Brazil
- **10. Apr 2021** Proposed EU AI Act (expires Q1 2024)
- **11. June 2021** South Korea Enforcement decree on Framework Act on Intelligent Informatization

2022

- **12. Mar 2022** China issues provisions on Internet Information Service Algorithm Recommendations and Administration of Deep Synthesis of Internet Information Services
- **13. June 2022** Canada's proposed Artificial Intelligence and Data Act (planned 2025)
- **14. Sept 2022** EU AI Liability Directive, a regime for dealing with damages caused by AI
- **15. Oct 2022** US Blueprint for an AI Bill of Rights
- **16. Dec 2022** Senate approval of the draft regulatory framework on artificial intelligence in Brazil

2023

- **17. Jan 2023** Stable Diffusion and Midjourney copyright lawsuits in the US
- **18. Jan 2023** NIST AI risk management framework
- **19. Feb 2023** South Korean Assembly proposed Act on Promotion of AI Industry and Framework for Establishing Trustworthy AI
- **20. Mar 2023** ChatGPT temporarily banned in Italy because of privacy concerns
- **21. Mar–Apr 2023** Several data protection regulators globally looking into ChatGPT data protection practices, eg, Germany, France, and Spain
- **22. Apr 2023** China released Draft Administrative Measures for Generative Artificial Intelligence Services
- **23. May 2023** Proposal for legal framework for artificial intelligence in Brazil merging previous proposals from 2019–21
- **24. Oct 2023** US presidential executive order on AI
- **25. Nov 2023** AI summit in UK
- **26. Dec 2023** Political agreement on EU AI Act

Source: OECD; McKinsey analysis

McKinsey & Company

What does this mean for organizations?

Organizations may be tempted to wait to see what AI regulations emerge. But the time to act is now. Organizations may face large legal, reputational, organizational, and financial risks if they do not act swiftly. Several markets, including Italy, have already banned ChatGPT because of privacy concerns, copyright infringement lawsuits brought by multiple organizations and individuals, and defamation lawsuits.

More speed bumps are likely. As the negative effects of AI become more widely known and publicized, public concerns increase. This, in turn, has led to public distrust of the companies creating or using AI.

A misstep at this stage could also be costly. Organizations could face fines from legal enforcement—of up to 7 percent of annual global revenues, according to the AI regulation proposed by the European Union, for example. Another threat is financial loss from falloff in customer or investor trust that could translate into a lower stock price, loss of customers, or slower customer acquisition. The incentive to move fast is heightened by the fact that if the right governance and organizational models for AI are not built early, remediation may become necessary later due to regulatory changes, data breaches, or cybersecurity incidents. Fixing a system after the fact can be both expensive and difficult to implement consistently across the organization.

The exact future of legal obligations is still unclear and may differ across geographies and depend on the specific role AI will play within the value chain. Still, there are some no-regret moves for organizations, which can be implemented today to get ahead of looming legal changes.

These preemptive actions can be grouped into four key areas that stem from existing data protection or privacy and cyber efforts, as they share a great deal of common ground:

Transparency. Create a taxonomy and inventory of models, classifying them in accordance with regulation, and record all usage across the organization in a central repository that is clear to those inside and outside the organization. Create detailed documentation of AI and gen AI usage, both internally and externally, its functioning, risks, and controls, and create clear documentation on how a model was developed, what risks it may have, and how it is intended to be used.

Governance. Implement a governance structure for AI and gen AI that ensures sufficient oversight, authority, and accountability both within the organization and with third parties and regulators. This approach should include a definition of all roles and responsibilities in AI and gen AI management and the development of an incident management plan to address any issues that may arise from AI and gen AI use. The governance structure should be robust enough to withstand changes in personnel and time but also agile enough to adapt to evolving technology, business priorities, and regulatory requirements.

Organizations are challenged with navigating varied regulations while planning their long-term AI strategies.

Data, model, and technology management. AI and gen AI both require robust data, model, and technology management:

— *Data management.* Data is the foundation of all AI and gen AI models. The quality of the data input also mirrors the final output of the model. Proper and reliable data management includes awareness of data sources, data classification, data quality and lineage, intellectual property, and privacy management.

— *Model management.* Organizations can establish robust principles and guardrails for AI and gen AI development and use them to minimize the organization's risks and ensure that all AI and gen AI models uphold fairness and bias controls, proper functioning, transparency, clarity, and enablement of human oversight. Train the entire organization on the proper use and development of AI and gen AI to ensure risks are minimized. Develop the organization's risk taxonomy and risk framework to include the risks associated with gen AI. Establish roles and responsibilities in risk management and establish risk assessments and controls, with proper testing and monitoring mechanisms to monitor and resolve AI and gen AI risks. Both data and model management require agile and iterative processes and should not be treated as simple tick-the-box exercises at the beginning of development projects.

— *Cybersecurity and technology management.* Establish strong cybersecurity and technology, including access control, firewalls, logs, monitoring, et cetera, to ensure a secure technology environment, where unauthorized access or misuse is prevented and potential incidents are identified early.

Individual rights. Educate users: make them aware that they are interacting with an AI system, and provide clear instructions for use. This should include establishing a point of contact that provides transparency and enables users to exercise their rights, such as how to access data, how models work, and how to opt out. Finally, take a customer-centric approach to designing and using AI, one that considers the ethical implications of the data used and its potential impact on customers. Since not everything legal is necessarily ethical, it is important to prioritize the ethical considerations of AI usage.

AI and gen AI will continue to have a significant impact on many organizations, whether they are providers of AI models or users of AI systems. Despite the rapidly changing regulatory landscape, which is not yet aligned across geographies and sectors and may feel unpredictable, there are tangible benefits for organizations that improve how they provide and use AI now.

Andreas Kremer is a partner in McKinsey's Berlin office; **Angela Luget** is a partner in the London office, where **Daniel Mikkelsen** is a senior partner; **Henning Soller** is a partner in the Franfurt office; **Malin Strandell-Jansson** is a senior knowledge expert in the Stockholm office; and **Sheila Zingg** is a consultant in the Zurich office.

The authors wish to thank Rachel Lee, Chris Schmitz, and Angie Selzer for their contributions to this article.

This article was edited by David Weidner, a senior editor in the Bay Area office.

4

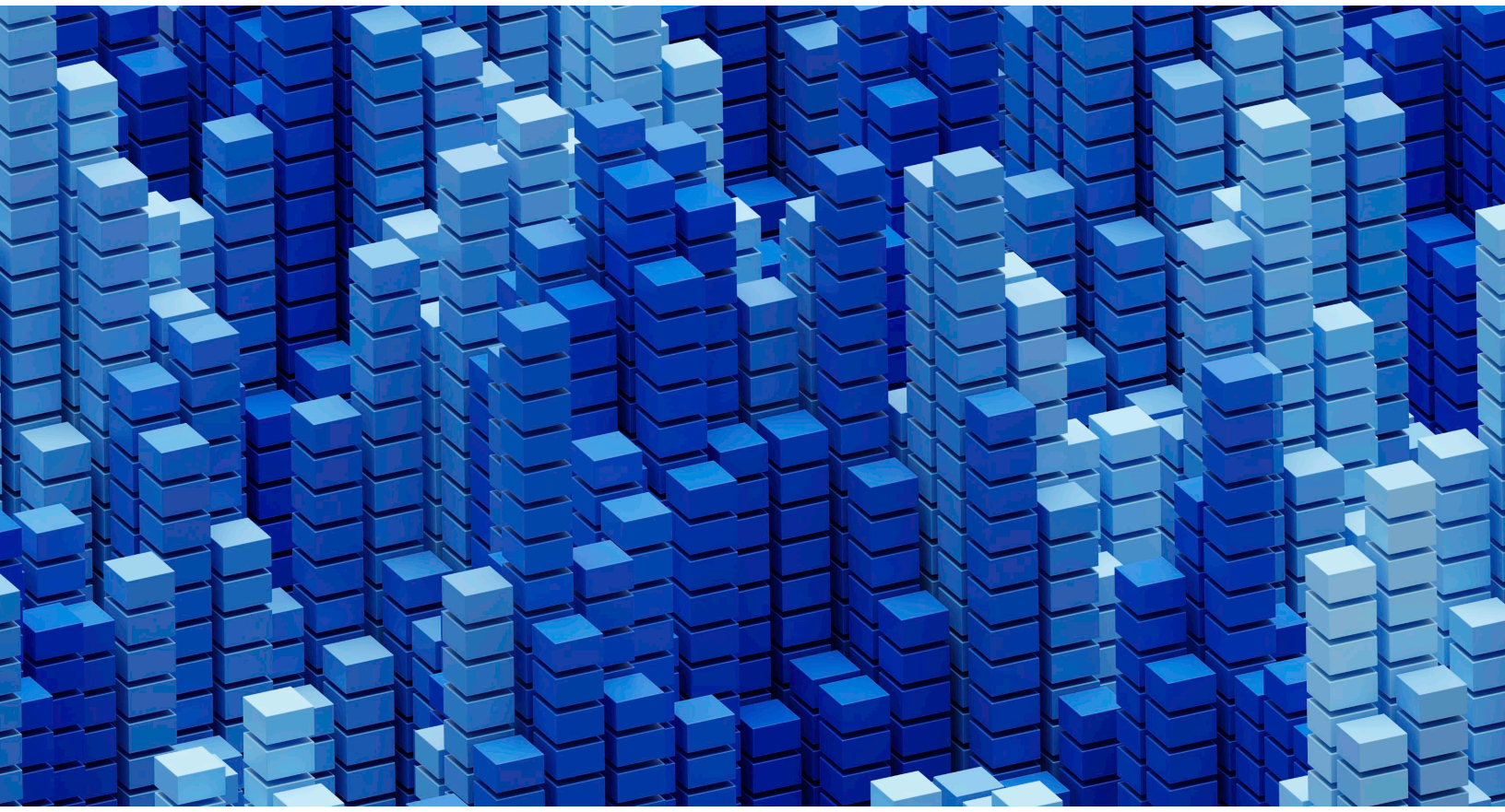
What the future holds

QuantumBlack
AI by McKinsey

Six major gen AI trends that will shape 2024's agenda

What every CEO needs to know and what it means for their gen AI agendas.

by Sam Bourton, Ben Ellencweig, Carlo Giovine, and Stephen Xu



November 2023

It's hard to believe that ChatGPT is only a year old. The number of exciting new product launches over the past 12 months has been astonishing—and there's no sign that things will slow down anytime soon. In fact, quite the opposite. Earlier in November, OpenAI hosted DevDay, where the company announced extensive offerings across B2C and B2B markets. Cohere has doubled down on its knowledge search capabilities and private deployments. And Amazon Web Services launched PartyRock, its no-code gen AI app-building playground.

We believe that this past month's activity is setting the stage for what to expect in 2024 in the gen AI space. Here are six major trends happening across the space.

1. **Gen AI can see, hear, and talk.** Multimodal models spanning text, code, image, and audio unlock new capabilities across both content generation and comprehension.

2. **Gen AI can interact with the world.** Gen AI models connect to data and IT systems to read and write data.

3. **Gen AI models are easier to control.** End users get more consistent outputs from probabilistic models via new features (for example, setting the seed).

4. **Gen AI development is being democratized.** OpenAI announced a new product called "GPTs," which allow nearly anyone to build a gen-AI-powered chatbot using low code/no code interfaces.

5. **Gen AI is a platform play.** Entire marketplaces of GPTs will be created. There, users will be able to discover new applications and publish their own.

6. **Gen AI costs continue to decline.** For one, GPT-4 API costs declined two to three times for the average enterprise customer.

While the technology's possibilities continue to grow, we believe there are four principles for CEOs to consider as they drive their gen AI agendas. The principles draw from our experiences building gen AI applications with our clients throughout 2023, as well as decades delivering digital and analytics transformations (covered thoroughly in our book, *Rewired*).

Be intentional. Set gen AI strategy top-down. Gen AI is a gold rush. Everyone from shareholders to employees to boards are scrambling to deploy the latest and most powerful gen AI tooling, and many large organizations have 150-plus gen AI use cases on backlog. While we share their excitement (and admire their ambition!), we have found that allowing dozens of gen AI projects to spawn across an organization puts at-scale value creation at risk.

With the recent developments in the gen AI space, the Cambrian explosion of the use cases and opportunities will only continue to split the already divided attention of leadership teams. C-suites must bring focus with a top-down gen AI strategy, while constantly returning to the question of how the technology can help create enduring strategic distance between the organization and its competitors. Here are some examples from first movers:

— Retail banks are increasing customer retention and offering conversion by deploying customer-facing chatbots and hyperpersonalization.

— Service operations are tackling ongoing labor shortages by building workflow copilots to improve productivity of existing labor forces to resolve customer requests on time.

— IT services players are growing market share by investing in software engineering productivity tools and pricing contracts more competitively.

Smart organizations are taking a 2x2 approach: identify two fast use cases to register quick wins on the scoreboard and excite the organization, while

working on two slower, more transformational use cases that will change day-to-day business operations.

Reimagine entire domains rather than isolated use cases.
During 2023, most organizations began experimenting with gen AI, building one-off prototypes and buying off-the-shelf solutions. Yet as these solutions are rolled out to end users, organizations are struggling to capture value. For example, some organizations that have invested in Github Copilot have yet to figure out how the value capture is passed back to the business. Organizations need to reframe from one isolated use case to the full software delivery life cycle. Scrum teams need to commit to shipping more product features. Or sales needs to offer more competitive pricing to their customers and win more business. If companies stop at just buying a new shiny tool, the productivity gains will not translate to bottom-line gains.

That often means reimagining entire workflows and domains. This serves two purposes: 1) it creates a more seamless end-user experience by avoiding point solutions; and 2) organizations can more easily track value against clear business outcomes. For example, an insurer we have worked with is reimagining its end-to-end claims process—from first notice of loss to payment. For each step along the way, the insurer has identified gen AI, digital, and analytics opportunities, while never losing sight of the claims adjustor's experience. Ultimately, this full sizing across the value chain made a step-change impact on end-to-end handling time.

Buy selectively. Build strategically.
Matching the pace of innovation, many new start-ups and software offerings are entering the market, leaving enterprises with a familiar question: "Buy or build?" On the "buy" side, we see organizations that are wary about investing in capabilities that likely will eventually be available for a fraction of the cost. These same organizations are also skeptical of off-the-shelf solutions, unsure if the software will perform at scale without significant customization. As these solutions mature and prove their value, buy strategies will continue to play a central role in any gen AI strategy.

Meanwhile, some organizations are finding compelling business cases to "build," as well. These players start by identifying use cases that create strategic competitive advantages against their peers, by compounding existing strengths in their domain expertise, workflow integration, or regulatory know-how. For example, deploying gen AI to accelerate drug discovery has become standard in the pharmaceutical industry. Additionally, organizations are making investments in data and IT infrastructure to enable their portfolio of gen AI use cases. For many organizations, there has been little to no investment in unstructured data governance. Now is the time.

Build products, not POCs.
With the new tooling available, a talented engineer can build a proof of concept over a weekend. In some cases, this might be sufficient to serve an enterprise need (for example, a summarization chatbot). However, for most use cases in a large enterprise context, proofs of concepts are not sufficient. They do not scale well into production and their performance rapidly degrades without the appropriate engineering and experimentation. At OpenAI's Dev Day, engineers showed how hard it is to turn a POC into a production-grade product. At the start, a demo POC only achieved 45 percent accuracy for a retrieval task. After a few months and a dozen or so experiments (for example, fine-tuning, reranking, metadata tagging, data labeling, model self-assessment, risk guardrails), the engineers achieved 98 percent accuracy.

This leads to two implications. First, organizations cannot seek near-perfection on every use case. They need to be selective about when it is worthwhile to invest scarce engineering talent to develop high-performance gen AI applications. For some situations, 45 percent accuracy may be sufficient to deliver business benefit. Second, organizations need to scale their gen AI capabilities to meet their ambitions. Most organizations have identified hundreds of gen AI use cases. And so, organizations are turning to reusable code components to accelerate development. Dedicated engineers, often sitting in a center of excellence (COE), codify best practices into these

code components, which allows subsequent gen AI efforts to build off the lessons learned from the trailblazing ones. We have seen these components accelerate delivery by 25 to 50 percent.

Throughout the past year, there has been an endless stream of gen AI news and hype. The coming year will likely be similar—but with a growing focus on delivering real business value to justify the billions in investment. From large enterprises to pioneering start-ups, organizations need to form their strategies around the decades-old principles from digital and analytics transformations. Organizations that get these tried-and-tested learnings right will form lasting strategic advantages against their competitors, creating sticky customer experiences and gaining market share in a challenging macroeconomic environment.

If 2023 was a year of hype, then 2024 will be the year of lasting impact at scale.

This article was originally published November 29, 2023 on the QuantumBlack, AI by McKinsey Medium blog.

Sam Bourton is a partner in the Lyon office; **Ben Ellencweig** is a senior partner in McKinsey's Stamford office; **Carlo Giovine** is a partner in the London office; and **Stephen Xu** is director of product management for Quantum Black, AI by McKinsey-R&D in the Toronto office.

Appendix: Generative AI solutions in action

There's plenty of hype around generative AI. But does the technology actually deliver?

Evidence already exists to show that the answer is yes, unmistakably. The gen AI solutions, or tools, listed in the exhibit come from an extensive pool of work McKinsey has done over the past year. These solutions can help you understand ways gen AI might fit within your organization. If you are intrigued by the technology and want to explore gen AI applications further, please reach out and schedule a session with our experts. During these sessions, we can provide in-depth demonstrations (either live or via video) of the technology and discuss how it can be tailored to meet your specific organizational needs *(for more information, visit McKinsey.com/GenAI).*

The gen AI landscape is evolving quickly. Don't miss out on the opportunity to stay ahead of the game.

Exhibit

Early solutions demonstrate the practical potential of generative AI.

Gen AI solution/tool	Description
Tech services	Document Q&A solution for ticket resolution in app-managed services projects to assist with user support queries
Media call agent	Tool that helps a global tech and media company generate fast, high-quality knowledge searches for agents serving customers during calls
Billing and revenue assurance	Knowledge synthesis agent that pulls data from diverse sources to assist with revenue cycle processes
Insight synthesis agent	Gen AI solution that helps a global information services company generate quicker and more meaningful insights into consumer behavior
Virtual agent	Tool to automate web research tasks, insight generation, and interaction with third-party APIs
Virtual analyst	Interactive chatbot that helps users easily extract insights from their personal data
Agent copilot	Tool providing real-time agent support and personalized customer recommendations
AI voice analytics	Application that understands call reasons and derives actionable insights to reduce demand, improve routing, and boost agent performance
Call center coaching	Tool that identifies opportunities for coaching call center employees on hard and soft skills

Gen AI solution/tool	Description
Voice-to-voice automation	Voice bot that responds to customer questions with customized answers and/or quick automated tasks
Marketing adviser	Tool that summarizes social media posts for marketing insights
Sales.ai	End-to-end solution leveraging analytical and gen AI to identify leads and conduct customer outreach at scale
Ada	One-point solution for gen AI procurement use cases (eg, negotiation preparation, contract review, idea generation, category facts, market and supplier news)
Doc IQ	Tool that uses advanced deep learning techniques to analyze and extract data from contract documents and images and collect the data into a structured format
RFP generator	Tool that lets users input the context of a request for proposal (RFP) and follow up with chat prompts to generate new RFP content automatically
Procurement contract AI	Gen AI tool that reviews contracts, comparing terms and clauses against predefined best practices and McKinsey proprietary knowledge, providing instant insights
Coding copilot	AI assistant to generate and troubleshoot code

Glossary

Application programming interface (API) is a way to programmatically access (usually external) models, datasets, or other pieces of software.

Artificial intelligence (AI) is the ability of software to perform tasks that traditionally require human intelligence.

Artificial neural networks (ANNs) are composed of interconnected layers of software-based calculators known as "neurons." These networks can absorb vast amounts of input data and process that data through multiple layers that extract and learn the data's features.

Deep learning is a subset of machine learning that uses deep neural networks, which are layers of connected "neurons" whose connections have parameters or weights that can be trained. It is especially effective at learning from unstructured data such as images, text, and audio.

Early and late scenarios are the extreme scenarios of our work-automation model. The "earliest" scenario flexes all parameters to the extremes of plausible assumptions, resulting in faster automation development and adoption; the "latest" scenario flexes all parameters in the opposite direction. The reality is likely to fall somewhere between the two.

Fine-tuning is the process of adapting a pretrained foundation model to perform better in a specific task. This entails a relatively short period of training on a labeled dataset, which is much smaller than the dataset the model was initially trained on. This additional training allows the model to learn and adapt to the nuances, terminology, and specific patterns found in the smaller dataset.

Foundation models (FMs) are deep learning models trained on vast quantities of unstructured, unlabeled data that can be used for a wide range of tasks out of the box or adapted to specific tasks through fine-tuning. Examples of these models are GPT-4, PaLM, DALL·E 2, and Stable Diffusion.

Generative AI is AI that is typically built using foundation models and has capabilities that earlier AI did not have, such as the ability to generate content. Foundation models can also be used for nongenerative purposes (for example, classifying user sentiment as negative or positive based on call transcripts) while offering significant improvement over earlier models. For simplicity, when we refer to generative AI, we include all foundation model use cases.

Graphics processing units (GPUs) are computer chips that were originally developed for producing computer graphics (such as for video games) and are also useful for deep learning applications. In contrast, traditional machine learning and other analyses usually run on central processing units (CPUs), normally referred to as a computer's "processor."

Large language models (LLMs) make up a class of foundation models that can process massive amounts of unstructured text and learn the relationships between words or portions of words, known as tokens. This enables LLMs to generate natural-language text, performing

tasks such as summarization or knowledge extraction. GPT-4 (which underlies ChatGPT) and LaMDA (the model behind Bard) are examples of LLMs.

Machine learning (ML) is a subset of AI in which a model gains capabilities after it is trained on, or shown, many example data points. Machine learning algorithms detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instruction. The algorithms also adapt and can become more effective in response to new data and experiences.

Modality is a high-level data category such as numbers, text, images, video, and audio.

Productivity from labor is the ratio of GDP to total hours worked in the economy. Labor productivity growth comes from increases in the amount of capital available to each worker, the education and experience of the workforce, and improvements in technology.

Prompt engineering refers to the process of designing, refining, and optimizing input prompts to guide a generative AI model toward producing desired (that is, accurate) outputs.

Self-attention, sometimes called intra-attention, is a mechanism that aims to mimic cognitive attention, relating different positions of a single sequence to compute a representation of the sequence.

Structured data is tabular data (for example, organized in tables, databases, or spreadsheets) that can be used to train some machine learning models effectively.

Transformers are a relatively new neural network architecture that relies on self-attention mechanisms to transform a sequence of inputs into a sequence of outputs while focusing its attention on important parts of the context around the inputs. Transformers do not rely on convolutions or recurrent neural networks.

Technical automation potential refers to the share of the work time that could be automated. We assessed the technical potential for automation across the global economy through an analysis of the component activities of each occupation. We used databases published by institutions including the World Bank and the US Bureau of Labor Statistics to break down about 850 occupations into approximately 2,100 activities, and we determined the performance capabilities needed for each activity based on how humans currently perform them.

Use cases are applications targeted to a specific business challenge that produce one or more measurable outcomes. For example, in marketing, generative AI could be used to generate creative content such as personalized emails.

Unstructured data lacks a consistent format or structure (for example, text, images, and audio files) and typically requires more advanced techniques to extract insights.

McKinsey & Company
February 2024

www.McKinsey.com

@McKinsey
@McKinsey
@McKinsey