McKinsey
& Company

QuantumBlack
AI by McKinsey

McKinsey Digital Practice

Why agents are the next frontier of generative AI

By moving from information to action—think virtual coworkers able to complete complex workflows—the technology promises a new wave of productivity and innovation.

By Lareina Yee, Michael Chui, and Roger Roberts
with Stephen Xu

July 2024

Over the past couple of years, the world has marveled at the capabilities and possibilities unleashed by generative AI (gen AI). Foundation models such as large language models (LLMs) can perform impressive feats, extracting insights and generating content across numerous mediums, such as text, audio, images, and video. But the next stage of gen AI is likely to be more transformative.

We are beginning an evolution from knowledge-based, gen AI–powered tools—say, chatbots that answer questions and generate content—to gen AI–enabled "agents" that use foundation models to execute complex, multistep workflows across a digital world. In short, the technology is moving from thought to action.

Broadly speaking, "agentic" systems refer to digital systems that can independently interact in a dynamic world. While versions of these software systems have existed for years, the natural-language capabilities of gen AI unveil new possibilities, enabling systems that can plan their actions, use online tools to complete those tasks, collaborate with other agents and people, and learn to improve their performance. Gen AI agents eventually could act as skilled virtual coworkers, working with humans in a seamless and natural manner. A virtual assistant, for example, could plan and book a complex personalized travel itinerary, handling logistics across multiple travel platforms. Using everyday language, an engineer could describe a new software feature to a programmer agent, which would then code, test, iterate, and deploy the tool it helped create.

Agentic systems traditionally have been difficult to implement, requiring laborious, rule-based programming or highly specific training of machine-learning models. Gen AI changes that. When agentic systems are built using foundation models (which have been trained on extremely large and varied unstructured data sets) rather than predefined rules, they have the potential to adapt to different scenarios in the same way that LLMs can respond intelligibly to prompts on which they have not been explicitly trained. Furthermore, using natural language rather than programming code, a human user could direct a gen AI–enabled agent system to accomplish a complex workflow. A multiagent system could then interpret and organize this workflow into actionable tasks, assign work to specialized agents, execute these refined tasks using a digital ecosystem of tools, and collaborate with other agents and humans to iteratively improve the quality of its actions.

In this article, we explore the opportunities that the use of gen AI agents presents. Although the technology remains in its nascent phase and requires further technical development before it's ready for business deployment, it's quickly attracting attention. In the past year alone, Google, Microsoft, OpenAI, and others have invested in software libraries and frameworks to support agentic functionality. LLM-powered applications such as Microsoft Copilot, Amazon Q, and Google's upcoming Project Astra are shifting from being knowledge-based to becoming more action-based. Companies and research labs such as Adept, crewAI, and Imbue also are developing agent-based models and multiagent systems. Given the speed with which gen AI is developing, agents could become as commonplace as chatbots are today.

What value can agents bring to businesses?

The value that agents can unlock comes from their potential to automate a long tail of complex use cases characterized by highly variable inputs and outputs—use cases that have historically been difficult to address in a cost- or time-efficient manner. Something as simple as a business trip, for example, can involve numerous possible itineraries encompassing different airlines and flights, not to mention hotel rewards programs, restaurant reservations, and off-hours activities, all of which must be handled across different online platforms. While there have been efforts to automate parts of this process, much of it still must be done manually. This is in large part because the wide variation in potential inputs and outputs makes the process too complicated, costly, or time-intensive to automate.

Gen AI–enabled agents can ease the automation of complex and open-ended use cases in three important ways:

- *Agents can manage multiplicity.* Many business use cases and processes are characterized by a linear workflow, with a clear beginning and series of steps that lead to a specific resolution or outcome. This relative simplicity makes them easily codified and automated in rule-based systems. But rule-based systems often exhibit "brittleness"—that is, they break down when faced with situations not contemplated by the designers of the explicit rules. Many workflows, for example, are far less predictable, marked by unexpected twists and turns and a range of possible outcomes; these workflows require special handling and nuanced judgment that makes rules-based automation challenging. But gen AI agent systems, because they are based on foundation models, have the potential to handle a wide variety of less-likely situations for a given use case, adapting in real time to perform the specialized tasks required to bring a process to completion.

- *Agent systems can be directed with natural language.* Currently, to automate a use case, it first must be broken down into a series of rules and steps that can be codified. These steps are typically translated into computer code and integrated into software systems—an often costly and laborious process that requires significant technical expertise. Because agentic systems use natural language as a form of instruction, even complex workflows can be encoded more quickly and easily. What's more, the process can potentially be done by nontechnical employees, rather than software engineers. This makes it easier to integrate subject matter expertise, grants wider access to gen AI and AI tools, and eases collaboration between technical and nontechnical teams.

- *Agents can work with existing software tools and platforms.* In addition to analyzing and generating knowledge, agent systems can use tools and communicate across a broader digital ecosystem. For instance, an agent can be directed to work with software applications (such as plotting and charting tools), search the web for information, collect and compile human feedback, and even leverage additional foundation models. Digital-tool use is both a defining characteristic of agents (it's one way that they can act in the world) but also a way in which their gen AI capabilities can uniquely be brought to bear. Foundation models can learn how to interface with tools, whether through natural language or other interfaces. Without foundation models, these capabilities would require extensive manual efforts to integrate systems (for example, using extract, transform, and load tools) or tedious manual efforts to collate outputs from different software systems.

How gen AI–enabled agents could work

Agents can support high-complexity use cases across industries and business functions, particularly for workflows involving time-consuming tasks or requiring various specialized types of qualitative and quantitative analysis. Agents do this by recursively breaking down complex workflows and performing subtasks across specialized instructions and data sources to reach the desired goal. The process generally follows these four steps (Exhibit 1):

1. *User provides instruction:* A user interacts with the AI system by giving a natural-language prompt, much like one would instruct a trusted employee. The system identifies the intended use case, asking the user for additional clarification when required.

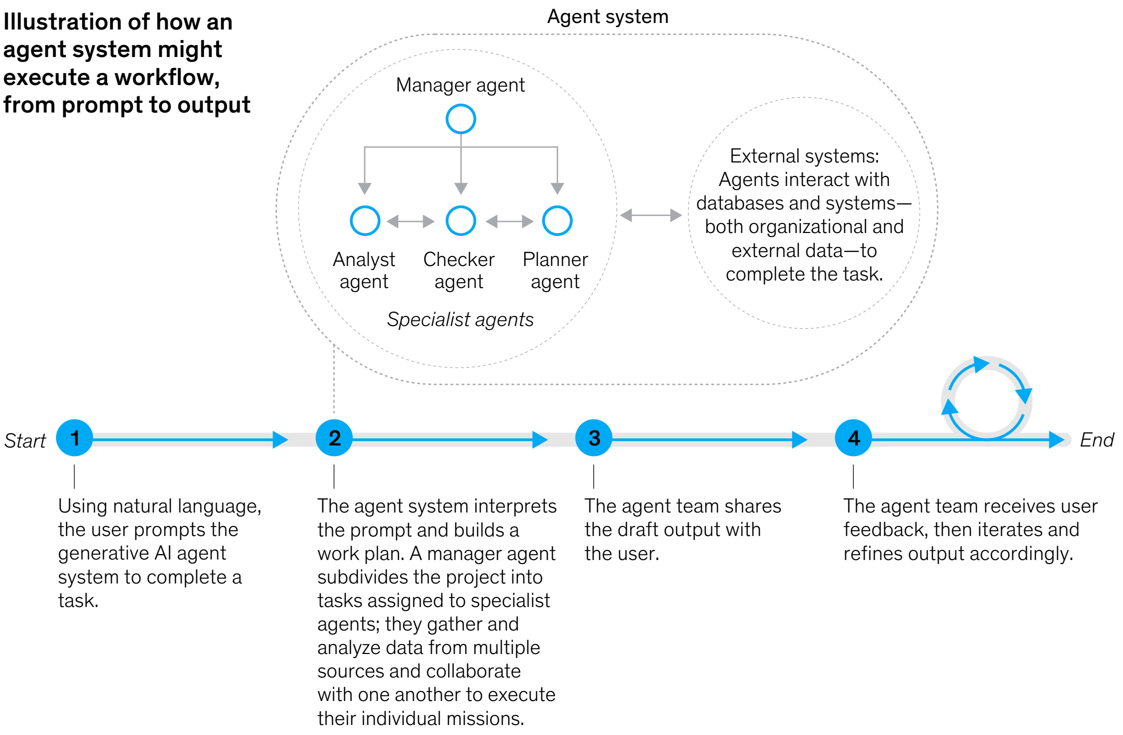
2. *Agent system plans, allocates, and executes work:* The agent system processes the prompt into a workflow, breaking it down into tasks and subtasks, which a manager subagent assigns to other specialized subagents. These subagents, equipped with necessary domain knowledge and tools, draw on prior "experiences" and codified domain expertise, coordinating with each other and using organizational data and systems to execute these assignments.

3. *Agent system iteratively improves output:* Throughout the process, the agent may request additional user input to ensure accuracy and relevance. The process may conclude with the agent providing final output to the user, iterating on any feedback shared by the user.

Exhibit 1

Agents enabled by generative AI soon could function as hyperefficient virtual coworkers.

Illustration of how an agent system might execute a workflow, from prompt to output


Agent system
Manager agent
Analyst agent
Checker agent
Planner agent
Specialist agents
External systems: Agents interact with databases and systems—both organizational and external data—to complete the task.
Start
1
2
3
4
End

Using natural language, the user prompts the generative AI agent system to complete a task.

The agent system interprets the prompt and builds a work plan. A manager agent subdivides the project into tasks assigned to specialist agents; they gather and analyze data from multiple sources and collaborate with one another to execute their individual missions.

The agent team shares the draft output with the user.

The agent team receives user feedback, then iterates and refines output accordingly.

McKinsey & Company

4. *Agent executes action:* The agent executes any necessary actions in the world to fully complete the user-requested task.

Art of the possible: Three potential use cases

What do these kinds of systems mean for businesses? The following three hypothetical use cases offer a glimpse of what could be possible in the not-too-distant future.

Use case 1: Loan underwriting

Financial institutions prepare credit-risk memos to assess the risks of extending credit or a loan to a borrower. The process involves compiling, analyzing, and reviewing various forms of information pertaining to the borrower, loan type, and other factors. Given the multiplicity of credit-risk scenarios and analyses required, this tends to be a time-consuming and highly collaborative effort, requiring a relationship manager to work with the borrower, stakeholders, and credit analysts to conduct specialized analyses, which are then submitted to a credit manager for review and additional expertise.

Potential agent-based solution: An agentic system—comprising multiple agents, each assuming a specialized, task-based role—could potentially be designed to handle a wide range of credit-risk scenarios. A human user would initiate the process by using natural language to provide a high-level work plan of tasks with specific rules, standards, and conditions. Then this team of agents would break down the work into executable subtasks.

One agent, for example, could act as the relationship manager to handle communications between the borrower and financial institutions. An executor agent could compile the necessary documents and forward them to a financial analyst agent that would, say, examine debt from cash flow statements and calculate relevant financial ratios, which would then be reviewed by a critic agent to identify discrepancies and errors and provide feedback. This process of breakdown, analysis, refinement, and review would be repeated until the final credit memo is completed (Exhibit 2).

Unlike simpler gen AI architectures, agents can produce high-quality content, reducing review cycle times by 20 to 60 percent. Agents are also able to traverse multiple systems and make sense of data pulled from multiple sources. Finally, agents can show their work: credit analysts can quickly drill into any generated text or numbers, accessing the complete chain of tasks and using data sources to produce the generated insights. This facilitates the rapid verification of outputs.

Use case 2: Code documentation and modernization

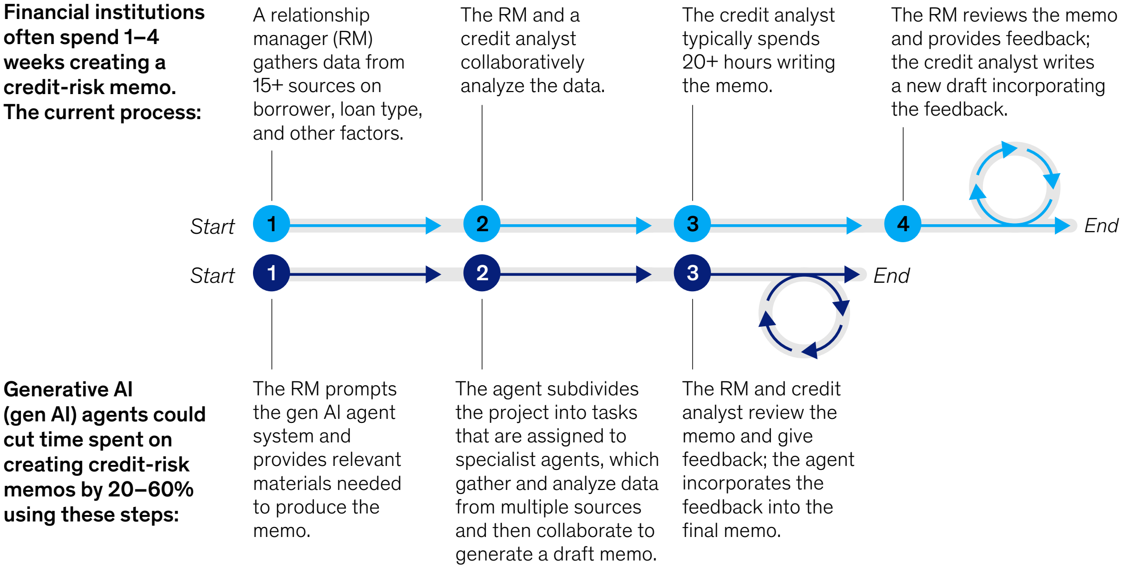
Legacy software applications and systems at large enterprises often pose security risks and can slow the pace of business innovation. But modernizing these systems can be complex, costly, and time-intensive, requiring engineers to review and understand millions of lines of the older codebase and manual documentation of business logic, and then translating this logic to an updated codebase and integrating it with other systems.

Potential agent-based solution: AI agents have the potential to significantly streamline this process. A specialized agent could be deployed as a legacy-software expert, analyzing old code and documenting and translating various code segments. Concurrently, a quality assurance agent could critique this documentation and produce test cases, helping the AI system to iteratively refine its output and ensure its accuracy and adherence to

Exhibit 2

Generative AI agents have the potential to change the way we work by supercharging productivity.

Illustrative use case: credit-risk memos



Financial institutions often spend 1–4 weeks creating a credit-risk memo. The current process:

1. A relationship manager (RM) gathers data from 15+ sources on borrower, loan type, and other factors.
2. The RM and a credit analyst collaboratively analyze the data.
3. The credit analyst typically spends 20+ hours writing the memo.
4. The RM reviews the memo and provides feedback; the credit analyst writes a new draft incorporating the feedback.

Generative AI (gen AI) agents could cut time spent on creating credit-risk memos by 20–60% using these steps:

1. The RM prompts the gen AI agent system and provides relevant materials needed to produce the memo.
2. The agent subdivides the project into tasks that are assigned to specialist agents, which gather and analyze data from multiple sources and then collaborate to generate a draft memo.
3. The RM and credit analyst review the memo and give feedback; the agent incorporates the feedback into the final memo.

McKinsey & Company

organizational standards. The repeatable nature of this process, meanwhile, could produce a flywheel effect, in which components of the agent framework are reused for other software migrations across the organization, significantly improving productivity and reducing the overall cost in software development.

Use case 3: Online marketing campaign creation
Designing, launching, and running an online marketing campaign tends to involve an array of different software tools, applications, and platforms. And the workflow for an online marketing campaign is highly complex. Business objectives and market trends must be translated into creative campaign ideas. Written and visual material must be created and customized for different segments and geographies. Campaigns must be tested with user groups across various platforms. To accomplish these tasks, marketing teams often use different forms of software and must move outputs from one tool to another, which is often tedious and time-consuming.

Potential agent-based solution: Agents can help connect this digital marketing ecosystem. For example, a marketer could describe targeted users, initial ideas, intended channels, and other parameters in natural language. Then, an agent system—with assistance from marketing professionals—would help develop, test, and iterate different campaign ideas. A digital marketing strategy agent could tap online surveys, analytics from customer relationship management solutions, and other market research platforms aimed at gathering insights to craft strategies using multimodal foundation models. Agents for content marketing, copywriting, and design could then build tailored content, which a human evaluator would review for brand alignment. These agents would collaborate to iterate and refine outputs and align toward an approach that optimizes the campaign's impact while minimizing brand risk.

How should business leaders prepare for the age of agents?

Although agent technology is quite nascent, increasing investments in these tools could result in agentic systems achieving notable milestones and being deployed at scale over the next few years. As such, it is not too soon for business leaders to learn more about agents and consider whether some of their core processes or business imperatives can be accelerated with agentic systems and capabilities. This understanding can inform future road map planning or scenarios and help leaders stay at the edge of innovation readiness. Once those potential use cases have been identified, organizations can begin exploring the growing agent landscape, utilizing APIs, tool kits, and libraries (for example, Microsoft Autogen, Hugging Face, and LangChain) to start understanding what is relevant.

To prepare for the advent of agentic systems, organizations should consider these three factors, which will be key if such systems are to deliver on their potential:

— *Codification of relevant knowledge:* Implementing complex use cases will likely require organizations to define and document business processes into codified workflows that are then used to train agents. Likewise, organizations might consider how they can capture subject matter expertise, which will be used to instruct agents in natural language, thus streamlining complex processes.

— *Strategic tech planning:* Organizations will need to organize their data and IT systems to ensure that agent systems can interface effectively with existing infrastructure. That includes capturing user interactions for continuous feedback and creating the flexibility to integrate future technologies without disrupting existing operations.

— *Human-in-the-loop control mechanisms:* As gen AI agents begin interacting with the real world, control mechanisms are essential to balance autonomy and risk (see sidebar, "Understanding the unique risks posed by agentic systems"). Humans must validate outputs for accuracy, compliance, and fairness; work with subject matter experts to maintain and scale agent systems; and create a learning flywheel for ongoing improvement.Organizations should start considering under what conditions and how such human-in-the-loop mechanisms should be deployed.

Understanding the unique risks posed by agentic systems

Large language models (LLMs), as we now know, are prone to mistakes and hallucinations. Because agent systems process sequences of LLM-derived outputs, a hallucination within one of these outputs could have cascading effects if protections are not in place. Additionally, because agent systems are designed to operate with autonomy, business leaders must consider additional oversight mechanisms and guardrails. While it is difficult to fully anticipate all the risks that will be introduced with agents, here are some that should be considered.

Potentially harmful outputs
Large language models are not always accurate, sometimes providing incorrect information or performing actions with undesirable consequences. These risks are heightened as generative AI (gen AI) agents independently carry out tasks using digital tools and data in highly variable scenarios. For instance, an agent might approve a high-risk loan, leading to financial loss, or it may make an expensive, nonrefundable purchase for a customer.

Mitigation strategy: Organizations should implement robust accountability measures, clearly defining the responsibilities of both agents and humans while ensuring that agent outputs can be explained and understood. This could be accomplished by developing frameworks to manage agent autonomy (for example, limiting agent actions based on use case complexity) and ensuring human oversight (for example, verifying agent outputs before execution and conducting regular audits of agent decisions). Additionally, transparency and traceability mechanisms can help users understand the agent's decision making process to identify potentially fraught issues early.

Misuse of tools
With their ability to access tools and data, agents could be dangerous if intentionally misused. Agents, for example, could be used to develop vulnerable code, create convincing phishing scams, or hack sensitive information.

Mitigation strategy: For potentially high-risk scenarios, organizations should build in guardrails (for example, access controls, limits on agent actions) and create closed environments for agents (for instance, limit the agent's access to certain tools and data sources). Additionally, organizations should apply real-time monitoring of agent activities with automated alerts for suspicious behavior. Regular audits and compliance checks can ensure that guardrails remain effective and relevant.

Insufficient or excessive human–agent trust
Just as in relationships with human coworkers, interactions between humans and AI agents are based on trust. If users lack faith in agentic systems, they might scale back the human–agent interactions and information sharing that agentic systems require if they are to learn and improve. Conversely, as agents become more adept at emulating humanlike behavior, some users could place too much trust in them, ascribing to them human-level understanding and judgment. This can lead to users uncritically accepting recommendations or giving agents too much autonomy without sufficient oversight.

Mitigation strategy: Organizations can manage these issues by prioritizing the transparency of agent decision making, ensuring that users are trained in the responsible use of agents, and establishing a humans-in-the-loop process to manage agent behavior. Human oversight of agent processes is key to ensuring that users maintain a balanced perspective, critically evaluate agent performance, and retain final authority and accountability in agent actions. Furthermore, agent performance should be evaluated by tying agents' activities to concrete outcomes (for example, customer satisfaction, successful completion rates of tickets).

In addition to addressing these potential risks, organizations should consider the broader issues raised by gen AI agents:

— *Value alignment:* Because agents are akin to coworkers, their actions should embody organizational values. What values should agents embody in their decisions? How can agents be regularly evaluated and trained to align with those values?

— *Workforce shifts:* By completing tasks independently, agent systems stand to significantly alter the way work is accomplished, potentially allowing humans to focus more on higher-level tasks that require critical thinking and managerial skills. How will roles and responsibilities shift in each business function? How can employees be provided with retraining opportunities? Are there new collaboration models that can enhance cooperation between humans and AI agents?

— *Anthropomorphism:* As agents increasingly have humanlike capabilities, users could develop an overreliance on them or mistakenly believe that AI assistants are fully aligned with their own interests and values. To what extent should humanlike characteristics be incorporated into the design of agents? What processes can be created to enable real-time detection of potential harms in human–agent interactions?

McKinsey's most recent "State of AI" survey found that more than 72 percent of companies surveyed are deploying AI solutions, with a growing interest in gen AI. Given that activity, it would not be surprising to see companies begin to incorporate frontier technologies such as agents into their planning processes and future AI road maps. Agent-driven automation remains an exciting proposition, with the potential to revolutionize whole industries, bringing a new speed of action to work.

That said, the technology is still in its early stages, and there is much development required before its full capabilities can be realized. The increased complexity and autonomy of these systems pose a host of challenges and risks. And if deploying AI agents is akin to adding new workers to the team, just like their human team members, agents will require considerable testing, training, and coaching before they can be trusted to operate independently. But even in these earliest of days, it's not hard to envision the expansive opportunities this new generation of virtual colleagues could potentially unleash.

We are celebrating the 60th birthday of the McKinsey Quarterly *with a yearlong campaign featuring four issues on major themes related to the future of business and society, as well as related interactives, collections from the magazine's archives, and more. This article is part of the campaign's Future of Technology issue. Sign up for the* McKinsey Quarterly *alert list to be notified as soon as other new* Quarterly *articles are published.*

Lareina Yee is a senior partner in McKinsey's Bay Area office, where **Michael Chui** and **Roger Roberts** are partners; **Stephen Xu** is a senior director of project management in the Toronto office.

The authors wish to thank Aneri Shah, Arun Mittal, Henry Zhang, Kimberly Te, Mara Pometti, and Rickard Ström for their contributions to this article.

Designed by McKinsey Global Publishing