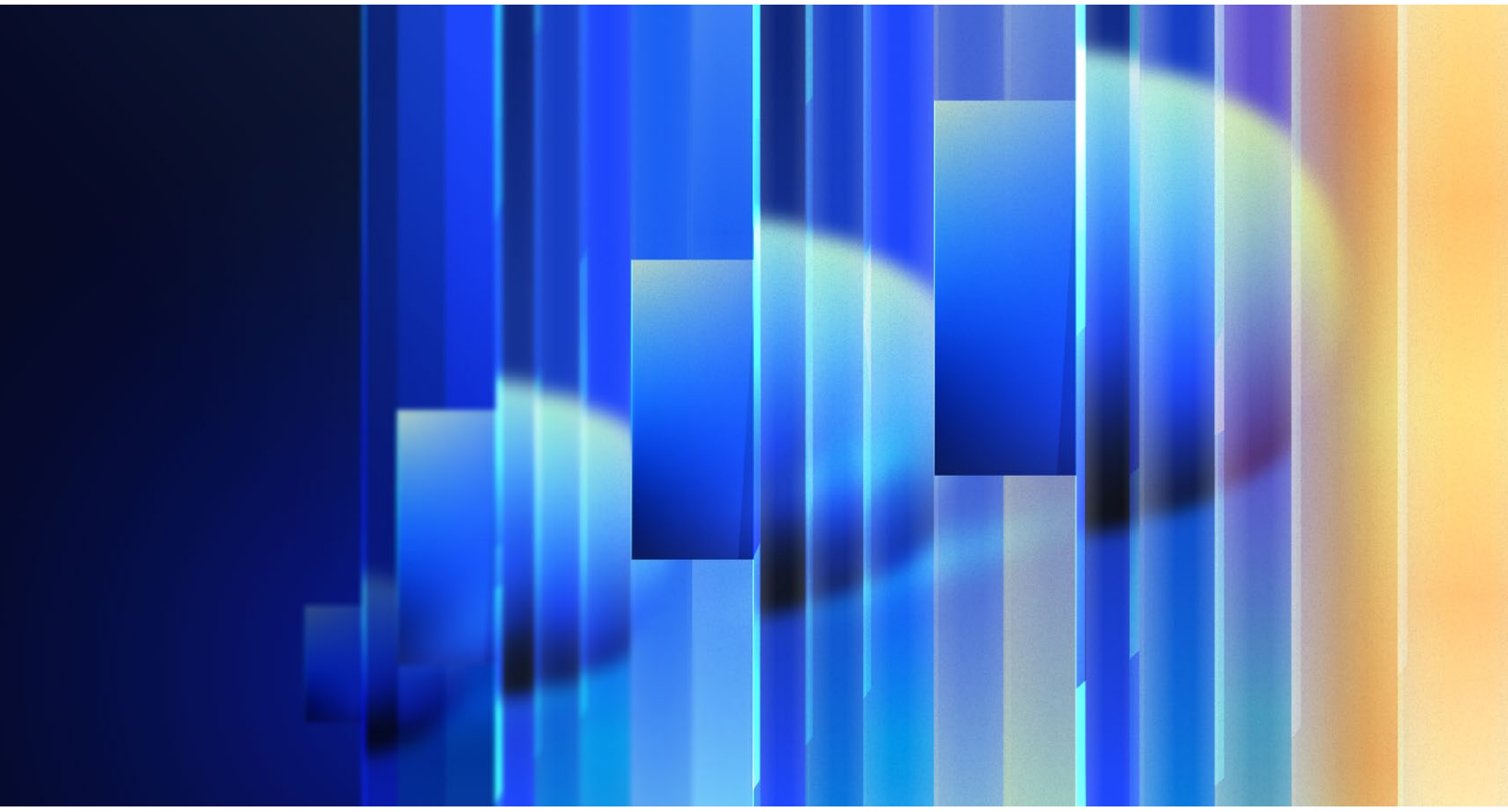
McKinsey & Company

Financial Services Practice

Extracting value from AI in banking: Rewiring the enterprise

To gain material value from AI, banks need to move beyond experimentation to transform critical business areas, including by reimagining complex workflows with multiagent systems.

This article is a collaborative effort by Carlo Giovine, Larry Lerner, Renny Thomas, Shwaitang Singh, Sudhakar Kakulavarapu, and Violet Chung, with Yuvika Motwani, representing views from McKinsey's Financial Services Practice.



December 2024

Much has been written about the power of AI, including generative AI (gen AI), to transform banking. Beyond ushering in the next wave of automation, AI promises to make banks more intelligent, efficient, and better able to achieve stronger financial performance.

While the buzz is undeniable, many banking C-suite leaders are increasingly asking questions about the realization of value in light of the headwinds facing the sector. Will AI live up to expectations? After initial experimentation, how can banks go from proof of concept to proof of value and truly reimagine and transform the enterprise using AI? How soon, if ever, can banks see a tangible return on their investments in AI?

These questions are gaining relevance as the global banking sector contends with challenges such as uneven labor productivity results, including falling productivity at US banks, despite high technology spending relative to other sectors. Banks also face slowing revenue and loan growth and competition from businesses beyond banking—such as private credit firms, fintechs, neobanks, payment solutions businesses, and nonbank providers—for the largest profit pools. To maintain the current return on tangible equity margins, banks will need to cut costs much faster as revenue growth slows. AI has the potential to chip away at these problems and put banks on more solid footing in the years to come, particularly in boosting labor productivity as employees continue to delegate a growing number of routine tasks to increasingly sophisticated and capable AI systems.

Some institutions are raising the bar and creating strategic distance from their peers by effectively scaling AI, including gen AI. For example, a large bank is using AI across the enterprise to improve experiences for its customers and employees, enhance efficiency, and boost revenue and profitability. In retail banking, the bank is harnessing AI to generate personalized nudges to help customers with investing and financial planning. In the small-business segment, AI is helping to pinpoint which loans might go bad, enabling the bank to take steps to intervene and support the client.

A regional bank, meanwhile, used gen AI to boost the productivity and efficiency of its software developers. Seeking to optimize resources and accelerate time to market of new developments, the bank launched a proof-of-concept study to assess the impact of gen AI tools on coding productivity. Productivity rose about 40 percent for the use cases that were part of the study; more than 80 percent of developers said gen AI improved their coding experience.

In this article, we detail a blueprint to help financial-services leaders chart the complex path of extracting at-scale value from AI across the enterprise. We begin with what banks that excel in AI do differently. We then outline a road map that roots the AI transformation in business value, ascertaining which key business problems need to be solved and harnessing technology, including AI, to help with the process. Next, we describe a comprehensive AI capability stack for banking powered by AI agents. Finally, we explore the elements needed to sustain and scale value from AI beyond the initial rollouts.

Delivering on the promise of AI in banking

The latest McKinsey Global Survey on AI shows that adoption has increased significantly across organizations and industries. However, the breadth of adoption (measured by the deployment of AI across multiple enterprise functions) remains low, and many organizations are still in the experimental phase.

Still, a few leading banks stand out in their ability to deploy AI, including gen AI, across the enterprise, and have begun to capture material gains from the use of AI (see sidebar "What does it mean to be an AI-first bank?").

What does it mean to be an AI-first bank?

AI is enabling broad changes in all sorts of industries, including banking, but many banks are still in the experimental phase. Given how far AI has come and the promise it holds, experimenting is not enough. To thrive in this new world, banks will need to become AI-first institutions, adopting AI technologies enterprise-wide to boost value—or risk being left behind.

A successful AI transformation spans several layers of the organization. It's important to invest in each of the interdependent layers, as underinvestment in one section can sabotage the entire AI transformation.

The essentials of building an AI-first bank include the following:

- Reimagining the customer experience by providing personalized offers and streamlined, frictionless use across various devices, for bank-owned platforms as well as partner ecosystems.
- Using AI to help with decision making, significantly enhancing productivity by building the architecture required to generate real-time analytical insights and translating them into messages addressing precise customer needs.
- Modernizing core technology required for the backbone of the AI capability stack, including automated cloud provisioning, an application programming interface, and streamlined architecture to enable continuous, secure data exchange among various parts of the bank.
- Setting up a platform operating model that brings together the right talent, culture, and organizational design.

Our experience suggests that banks excelling in AI do four things well:

- *Set a bold, bankwide vision for the value AI can create.* Leading banks have an expansive outlook on the role that AI can play, viewing the technology not just as a driver of cost efficiencies but also as a way to enhance revenues and significantly improve customer and employee experiences.
- *Root the transformation in business value by transforming entire domains, processes, and journeys rather than just deploying narrow use cases.* Banks that excel in AI resist the temptation to launch narrow use cases such as a chatbot or a conversational Q&A tool in isolation. Although these might be fast to launch and potentially low risk, in isolation, they won't unlock material financial value.
- *Build a comprehensive stack of AI capabilities powered by multiagent systems.* Running complex banking workflows, such as evaluating a commercial customer's loan application, involves highly variable steps and the processing of a mix of structured and unstructured data. While traditional automation cannot handle such tasks, gen-AI-enabled multiagent systems, combined with predictive AI and digital tools, can (see sidebar "What are multiagent systems?"). Expanding these systems to the entire enterprise requires setting up a comprehensive AI bank stack.
- *Sustain and scale value by setting up critical enablers of the AI transformation.* These include cross-functional business, technology, and AI teams along with a central AI control tower that coordinates enterprise decisions across functions, drives governance and adoption of standardized risk guardrails, and promotes the reusability of AI capabilities.

Setting a bold, bankwide vision for the value AI can create

McKinsey's experience with hundreds of companies across various industries shows that capturing value from digital and AI transformations requires a fundamental rewiring of how a company operates. This involves six critical enterprise capabilities: a business-led digital road map, talent with the right skills, a fit-for-purpose operating model, technology that's easy for teams to use, data that's continually enriched and easily accessible across the enterprise, and adoption and scaling of digital solutions. These elements are interconnected, and all have to function well for the transformation to be a success.

What are multiagent systems?

Multiagent systems, also known as agentic systems, have been around for years but have been kicked into a higher gear in the past two years, thanks to the natural-language capabilities of generative AI (gen AI). Although they are still in a nascent phase, and much of the value they could generate remains hypothetical, multiagent systems are expected to improve over time.

These systems could be capable of planning actions, using tools to complete those actions, collaborating with other agents and people, and improving their performance as they learn by doing.

Eventually, gen AI agents could act as virtual coworkers. For instance, an engineer could use everyday language to describe a new software feature to a programmer agent, which would then code, test, iterate, and deploy the tool it helped create.

AI can do much more than just automate processes and boost efficiency. Banks that extract value from AI view the technology as a transformational tool and use AI for core strategic priorities such as boosting revenue, differentiating the bank from competitors, and driving higher satisfaction for customers and employees.

Leading banks embed AI in the strategic planning process, requiring every business unit to revamp its operations and set bold financial and customer goals. They focus on innovation by prioritizing the most high-impact areas that are core to strategy, versus experimenting in peripheral areas seen as safe bets or taking the "peanut butter" approach by spreading investments across many disparate initiatives. Next, they invest in enabling the scalability of AI initiatives by setting up the right data and technology platforms.

Leading banks also ensure that major AI initiatives are business led, not just technology led. This means business executives take ownership of shaping the design of interventions, ensuring what is built is tightly aligned with what the business needs, and holding joint accountability with technology leaders to deliver outcomes.

Rooting the transformation in business value

Launching a chatbot, creating a document summarizer, using off-the-shelf gen AI tools to create ads and write emails—although these types of AI endeavors allow banks to experiment and learn with minimal risks involved, the results are typically incremental and, in isolation, rarely lead to material changes in financial outcomes.

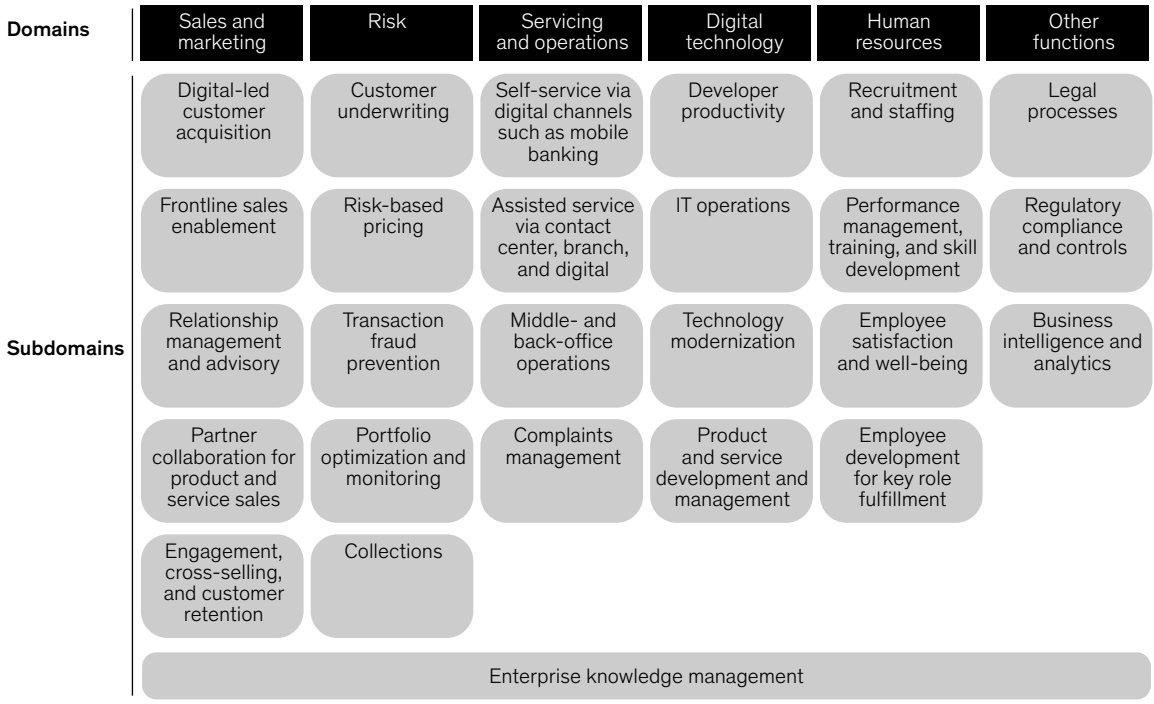
Using AI to significantly boost business value will require banks to do the following:

— *Choose the right scope of transformation by rewiring entire domains and subdomains.* Instead of letting a thousand flowers bloom with many disparate, siloed AI projects, leading banks are using AI to reimagine entire business domains—such as risk, sales, and operations—and within them, subdomains such as relationship management, collections, and contact-center servicing and operations. A typical bank has roughly 25 subdomains (Exhibit 1). Once bank executives choose the subdomains for transformation, they reimagine each one end to end, using the full range of AI and digital technologies to achieve the desired financial outcomes.

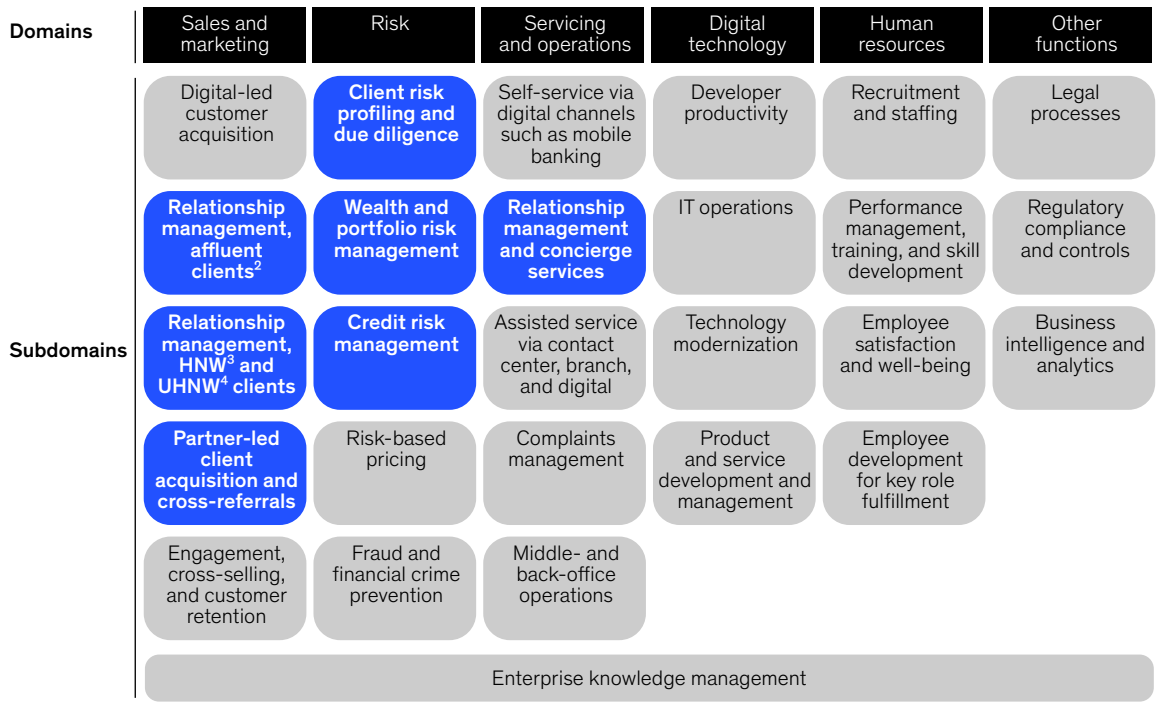
Exhibit 1

Banks can identify business areas for AI transformation and then rewire them to boost value.

Examples of subdomains that AI could transform in retail banking[1]



Examples of subdomains that AI could transform in private banking[1]



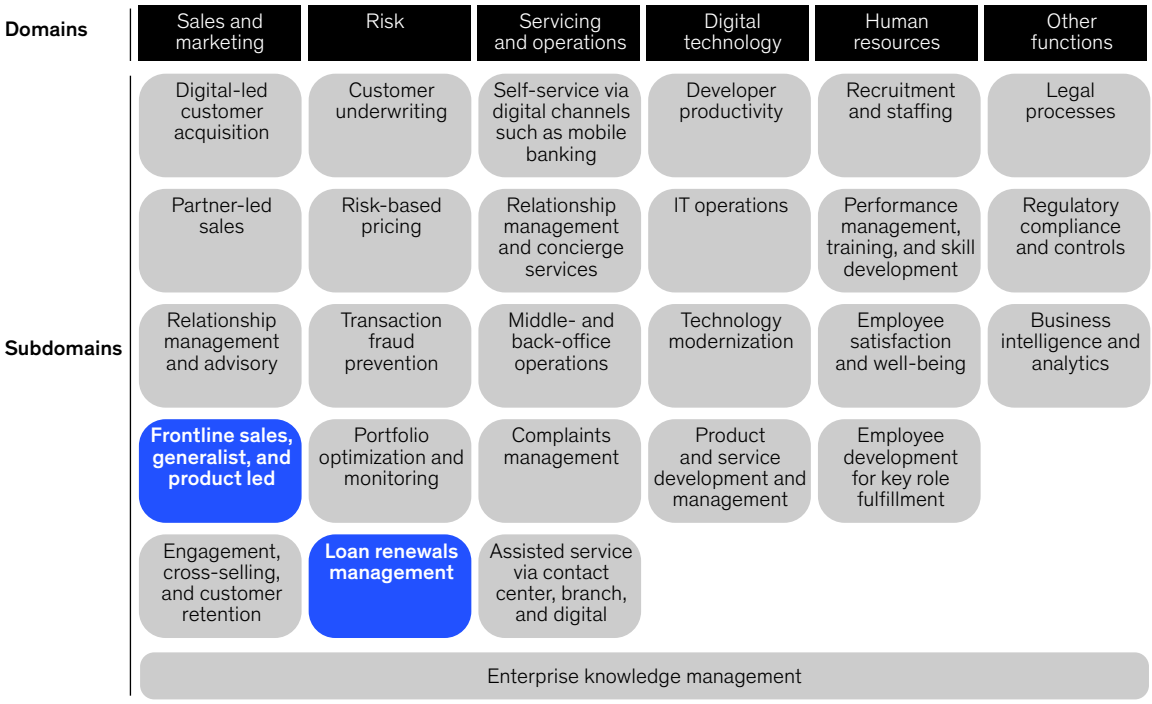
[1]A typical bank has ~25 subdomains that could be rewired with AI. This list is not comprehensive.
[2]Clients with personal financial assets of $100,000–$1 million.
[3]High-net-worth clients are those with personal financial assets of $1 million–$50 million.
[4]Ultra-high-net-worth clients are those with personal financial assets of >$50 million.

McKinsey & Company

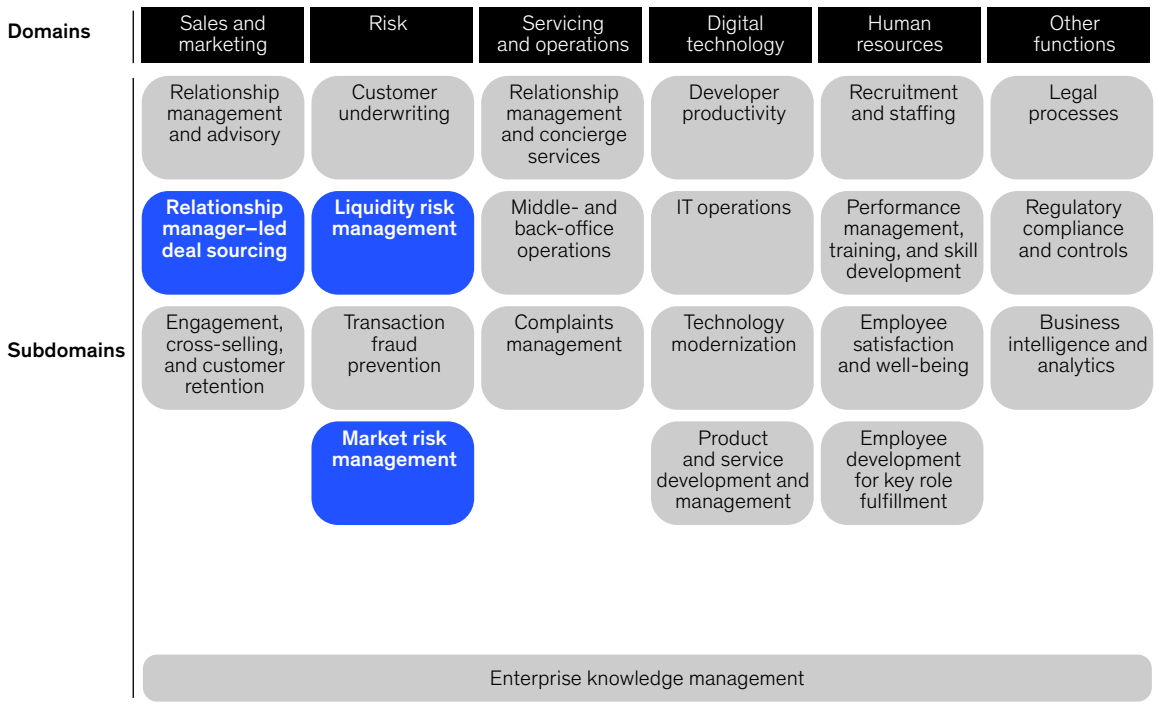
Exhibit 1 (continued)

Banks can identify business areas for AI transformation and then rewire them to boost value.

Examples of subdomains that AI could transform in corporate and commercial banking[1]



Examples of subdomains that AI could transform in investment banking[1]



[1]A typical bank has ~25 subdomains that could be rewired with AI. This list is not comprehensive

McKinsey & Company

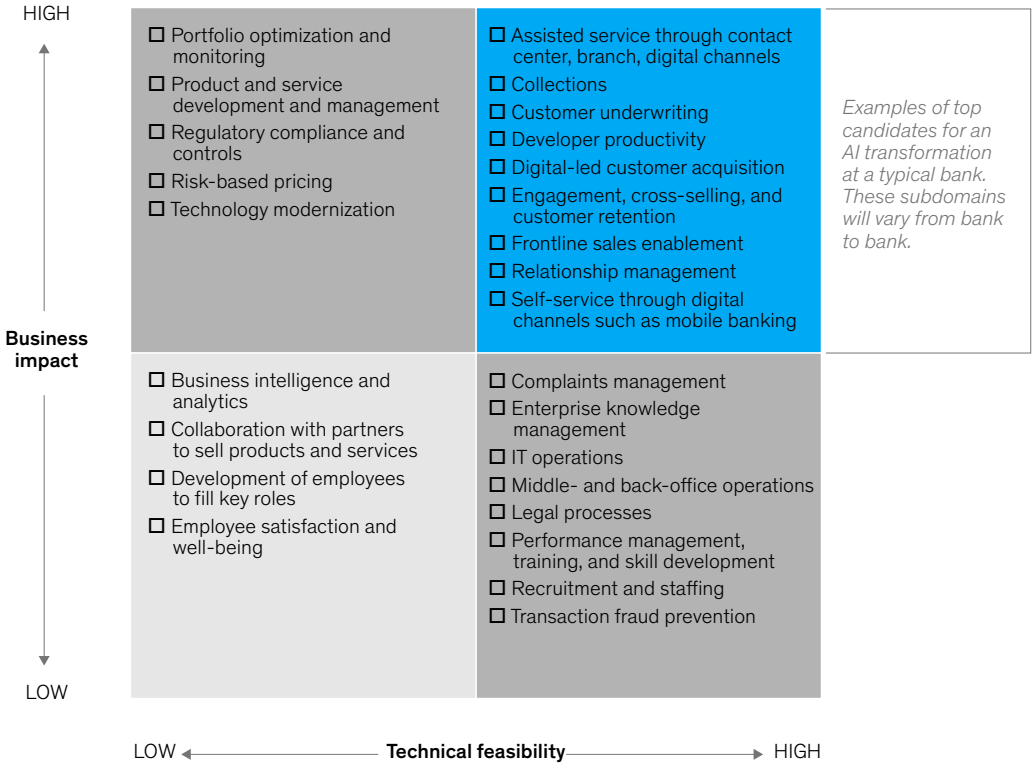
— *Decide which subdomains to transform with AI and in which order.* To select these subdomains, banks can consider the overall business impact and technical feasibility of driving an AI transformation of a specific subdomain (and the likelihood that the chosen subdomain includes components that can be reused in subsequent subdomain transformations) (Exhibit 2). In our experience, a typical bank has fewer than ten subdomains that could most benefit from an AI overhaul and should be the first candidates for transformation. Together, these subdomains can drive 70 to 80 percent of total incremental value from an AI transformation.

In terms of business impact, banks will need to assess whether the value of an AI transformation of a particular subdomain can be accurately quantified, how well the proposed solution aligns with the bank's strategic objectives, how well end users (whether clients or employees) are equipped to adopt the solution, and whether the solution will be a priority for the business.

Exhibit 2

Bank subdomains with high business impact and high technical feasibility should be first in line for an AI transformation.

Illustrative example of how business impact and technical feasibility can inform the transformation



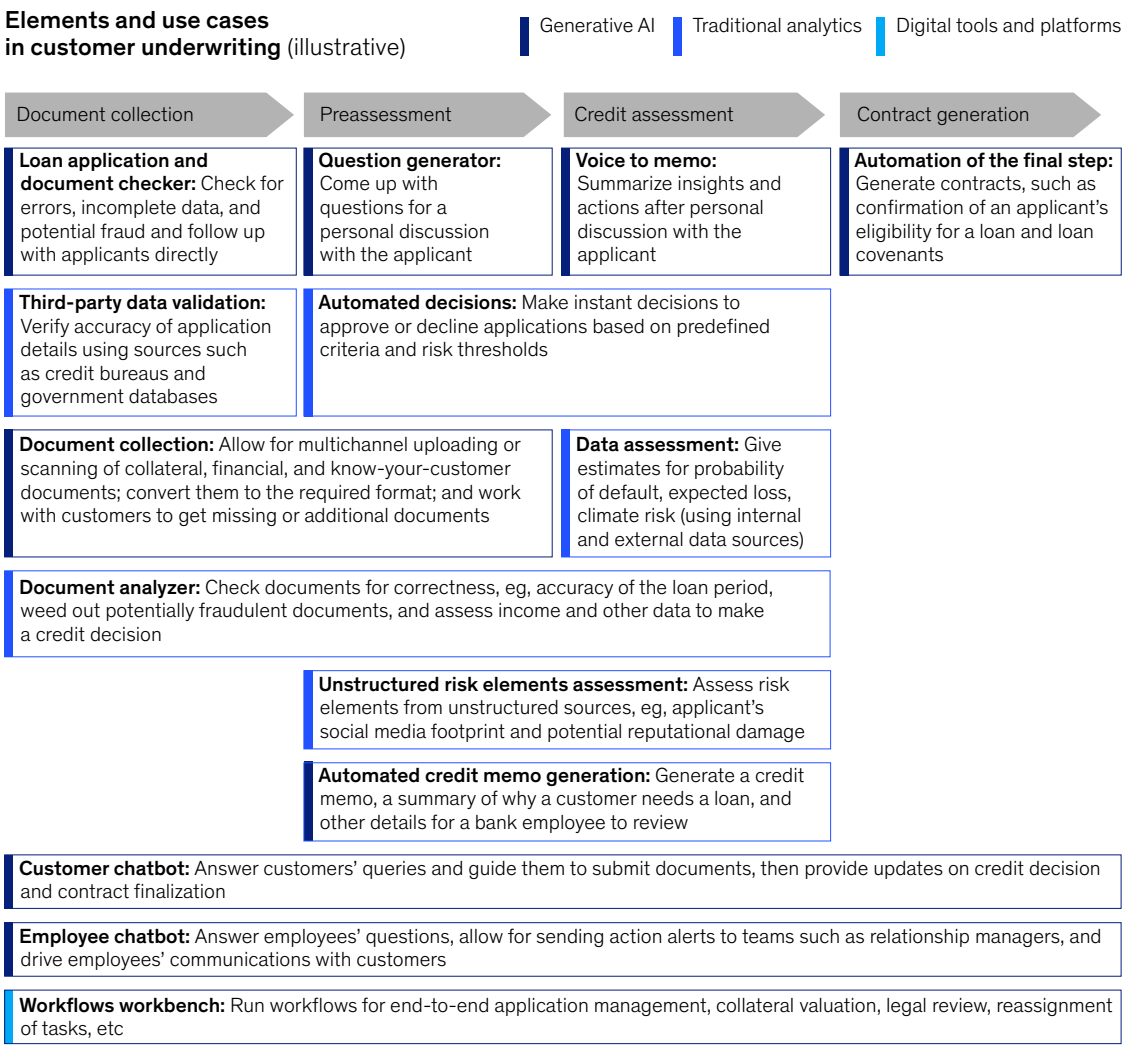
McKinsey & Company

Regarding technical feasibility, it is important to ascertain the availability and quality of data, including special considerations for handling sensitive data, techniques for scaling the solution across other domains and business units, the reusability of the solution's components for other use cases, and the presence of legacy technology infrastructure that may not be compatible with more modern AI solutions.

Once selected for an AI transformation, each subdomain can be deconstructed into a series of executable modules that need to be built, delivered, and adopted to drive business value. For example, transforming the customer underwriting subdomain end to end involves gen AI, traditional analytics, and digital tools and platforms all working together to reimagine end-to-end workflows and processes (Exhibit 3).

Exhibit 3

Banks can rewire the customer underwriting subdomain by using a combination of gen AI, traditional analytics, and digital tools and platforms.



McKinsey & Company

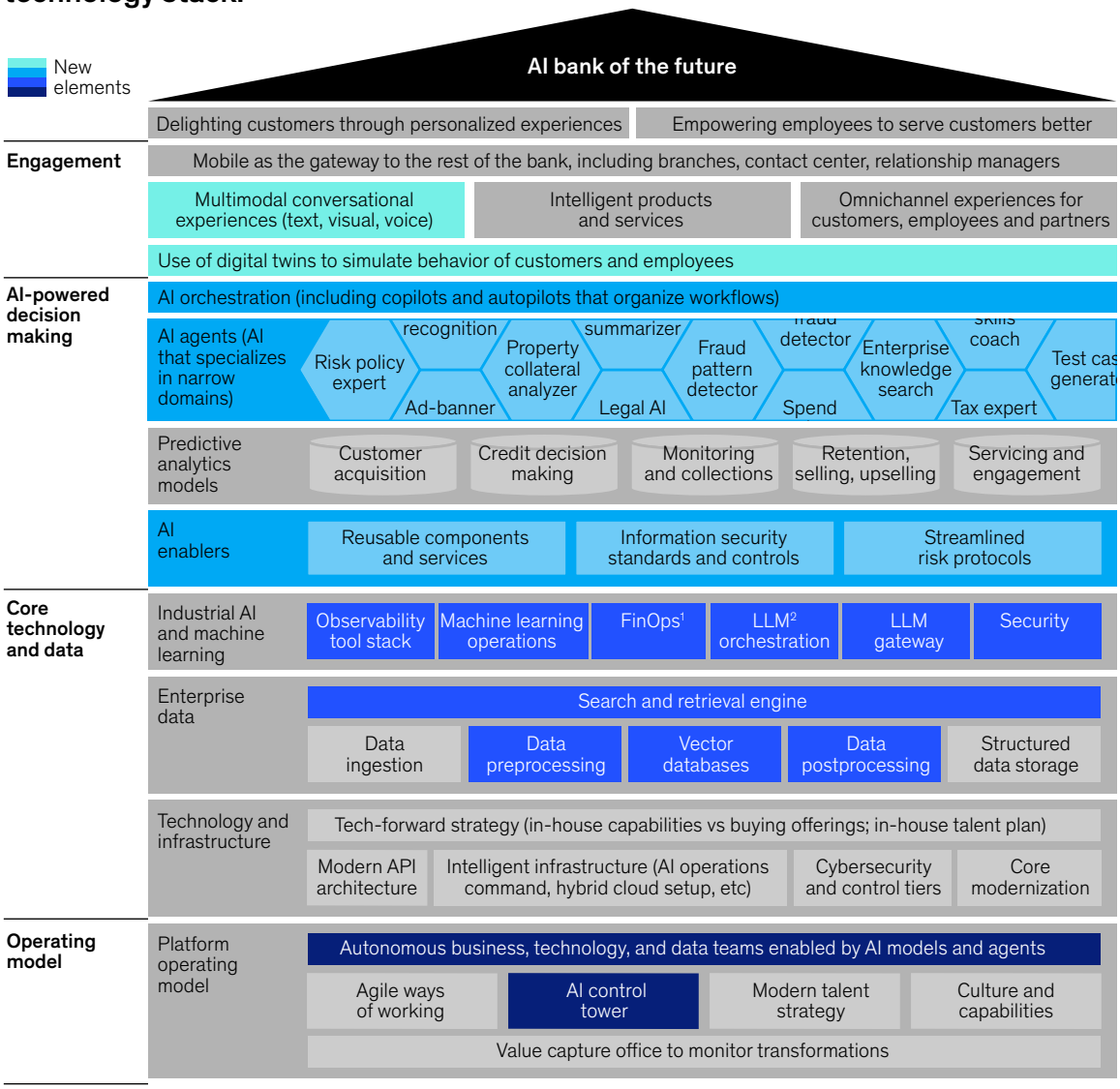
Enabling value through an AI stack powered by multiagent systems

To embed AI seamlessly across the enterprise, banks can implement a comprehensive capability stack that goes beyond just AI models. This AI bank stack contains four key capability layers: engagement, decision making, data and core tech, and operating model. Each layer will need to receive investment and attention to unlock the full power of AI for the enterprise.

Given the advent of new technologies such as gen AI, we have updated the AI capability stack (Exhibit 4) from a previous iteration published in 2020. Each layer's foundational elements are supplemented by several new elements.

Exhibit 4

To drive sustainable value, banks need to put AI first and revamp the entire technology stack.



[1]Financial operations, a framework for managing the operational costs of cloud computing.
[2]Large language models.

McKinsey & Company

The AI bank of the future
To create sustainable value, banks need to put AI first and revamp the entire technology stack. The rise of innovative technologies such as gen AI has prompted an update to the technology stack from a previous version published in 2020, with new elements highlighted in shades of blue.

Engagement layer
Banks will need to reimagine how they engage with customers, making their experiences as intelligent, personalized, and frictionless as possible through the use of AI. Leading banks' customers are experiencing human-like conversational interactions with AI via text and voice chats and are moving seamlessly across channels such as mobile apps, websites, branches, and contact centers, thanks to powerful AI capabilities.

AI-powered decision-making layer
The brain of the bank, this layer makes and orchestrates decisions. Historically, banks have focused on deploying traditional analytics modules such as models, but as AI technologies mature, this layer has expanded to include agent and AI orchestration sublayers working in unison with the traditional analytics layer to drive superior outcomes.

Core technology and data layer
This layer includes the technology and data needed for an AI transformation, including reusable tools and pipelines equipped with machine learning operations capabilities needed to run large language models (LLMs) at scale. Other portions of this layer include the data needed to train multiagent systems, as well as modern application programming interface (API) architecture and robust cybersecurity.

Operating model
By integrating business and technology in platforms run by cross-functional teams, banks can break up organizational silos, boost agility and speed, and better align goals and priorities across the enterprise. An AI control tower tracks the value realized from AI initiatives, among other tasks.

All together now
Elements across the four layers of the AI bank stack work together to enable transformative change and deliver value for the enterprise.

The key to next-generation innovation and productivity: Orchestrated multiagent systems
The decision-making layer is the brain of the AI-first bank, orchestrating and enabling thousands of AI-powered decisions affecting customers (such as which product to recommend to them next) and employees (for instance, should they approve credit for a specific customer or flag a transaction as fraudulent) across the full life cycle of products and services.

Predictive AI models, a core part of the decision-making layer at most banks, are great at driving decisions when presented with structured data under controlled conditions. These models, however, struggle to adapt when data is unstructured and the nature of the tasks is nonlinear and requires multistep planning, reasoning, and orchestration. Such tasks include, for example, preparing a credit memo—a summary of why a customer needs a loan and other details—based on multiple interactions with that customer and an evaluation of various types of documents. Another example is coaching a low-performing seller on how to improve sales performance.

Orchestrated multiagent systems represent a major advancement in the decision-making layer. These systems comprise various AI "agents" that can be thought of as virtual coworkers. Enabled by advances in gen AI technology, these agents, like humans, have the capacity to eventually be able to plan (for instance, organize a workflow encompassing a series of tasks), think (come up with chain-of-thought reasoning), and act (use digital tools).

Multiagent systems remain nascent and will need more technical development before they will be ready to deploy at scale across enterprises, but they are nonetheless attracting attention because of the promise they hold.

These agents, when combined with predictive AI models and digital tools, could fundamentally rewire several domains of the bank, not just unlocking productivity but forming the basis of more engaging experiences for customers and bank employees.

Multiagent systems can automate complex decisions and workflows through a twofold use of AI (Exhibit 5):

— *The AI orchestration layer handles complex workflows and task planning.* These AI orchestrators, programmed to achieve certain goals, are expected to eventually be able to do things like autonomously plan actions, reach decisions, and make use of existing tools, in-house data, and other AI agents to complete stated goals. These AI orchestrators could manifest themselves in the form of increasingly popular copilots for employees and customers. For instance, a copilot for a bank's credit managers is capable of not just answering questions but orchestrating the entire credit workflow when a customer applies for a loan. For now, human intervention is still needed to train and define operating procedures for the orchestrators to follow; however, the hope is that the technology will evolve to make them more autonomous.

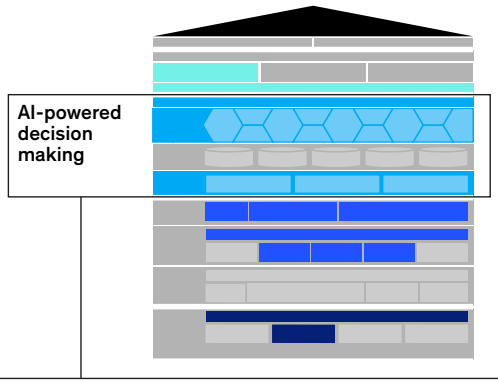
— *The AI agent layer comprises AI focused on completing specialized tasks as instructed by the orchestration layer or by other agents.* Each of these AI agents, powered by LLMs, is fine-tuned through a combination of domain-specific data and human feedback. For instance, a policy agent, after being provided with the bank's loan policies and related exceptions, can suggest the appropriate loan terms for a customer, much like a seasoned bank executive would. Meanwhile, a collateral inspection agent can be trained on documents and images related to collateral that customers use to apply for a loan, such as photos of small-business storefronts. A computer vision tool would then work with the collateral inspection agent to screen new collateral documents and images to spot instances of fraud, such as doctored photos that make the storefront look more impressive than it really is.

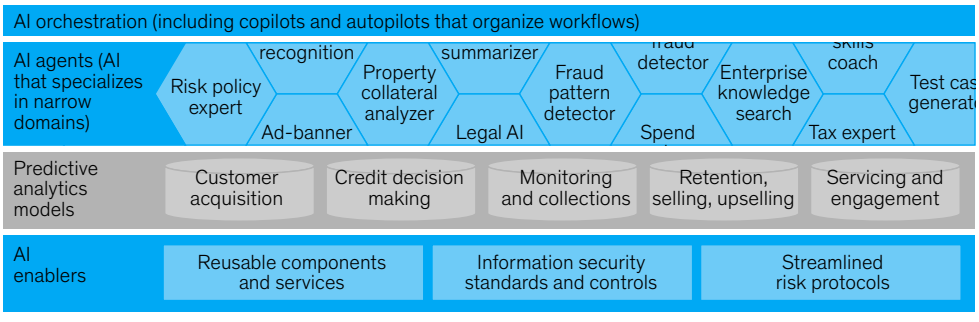
Exhibit 5

Orchestrated multiagent systems represent a big advance in banks' decision-making capabilities.

Interacting to solve problems

Multiagent systems are still in a nascent phase, but eventually AI agents could act as virtual coworkers capable of planning and executing tasks.





The AI orchestration layer handles complex workflows, calling on specialized AI agents to complete discrete tasks. Analytics and AI enablers support these efforts.

McKinsey & Company

The work of these two types of AI is enhanced by AI enablers, including reusable components and services encompassing tasks such as retrieval, intent detection, and summarization, as well as safety elements such as data confidentiality and enterprise-level risk guardrails. All of these put together ensure that models are deployed with speed and safety.

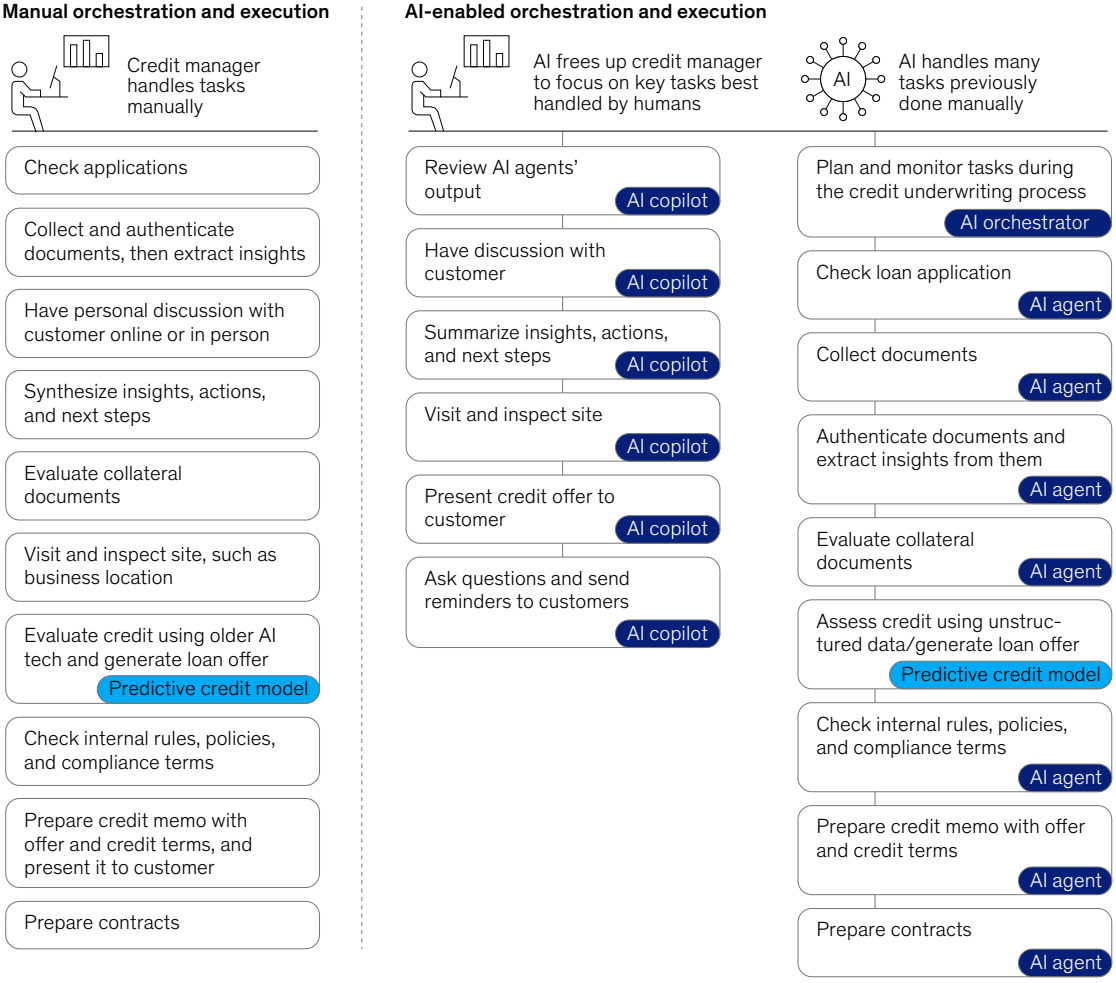
Over time, banks could have hundreds of AI agents at their disposal, each trained to complete a particular task and ready to be called on by other agents or humans. These agents can be continuously trained to become better over time, and they can be embedded across workflows. Humans will continue to oversee the agents, frequently auditing the results generated by multiagent systems and adjusting as needed.

As an illustration of how this works, consider how the traditionally complex task of underwriting credit for a small-business customer can be revamped through a mix of AI orchestrators and agents (Exhibit 6). The traditional way to do this is for humans to handle every step, moving from

Exhibit 6

The credit manager's role has the potential to evolve significantly thanks to AI.

How AI can help human credit managers achieve higher efficiency and effectiveness



McKinsey & Company

document collection to a discussion with the customer to assessment of collateral and so on. With orchestrated multiagent systems, agents can handle most of these tasks. A credit manager steps in to review the agents' output and handle tasks that require the human touch: chatting with the customer, visiting the small business in question, and the final step, presenting the credit offer to the customer.

As a result, credit risk teams could evolve into a mix of humans working with many AI orchestrators and agents.

When implemented well, multiagent systems can fundamentally rewire various domains at a bank. For example, we analyzed the effects of using multiagent systems to prepare credit memos and found credit analyst productivity gains of 20 to 60 percent, depending on various factors, and roughly 30 percent faster decision making.

Beyond boosting productivity, the use of multiagent systems can form the basis of more engaging experiences for customers and bank employees. For instance, a multiagent system can help customers during a loan application process even if they don't have all the required documents, enabling them to move on to the next step and ensuring that the documents are requested later. For employees, a multiagent system could help a sales associate who is underperforming by creating a conversational experience that could offer the employee specific actions to secure the next sale.

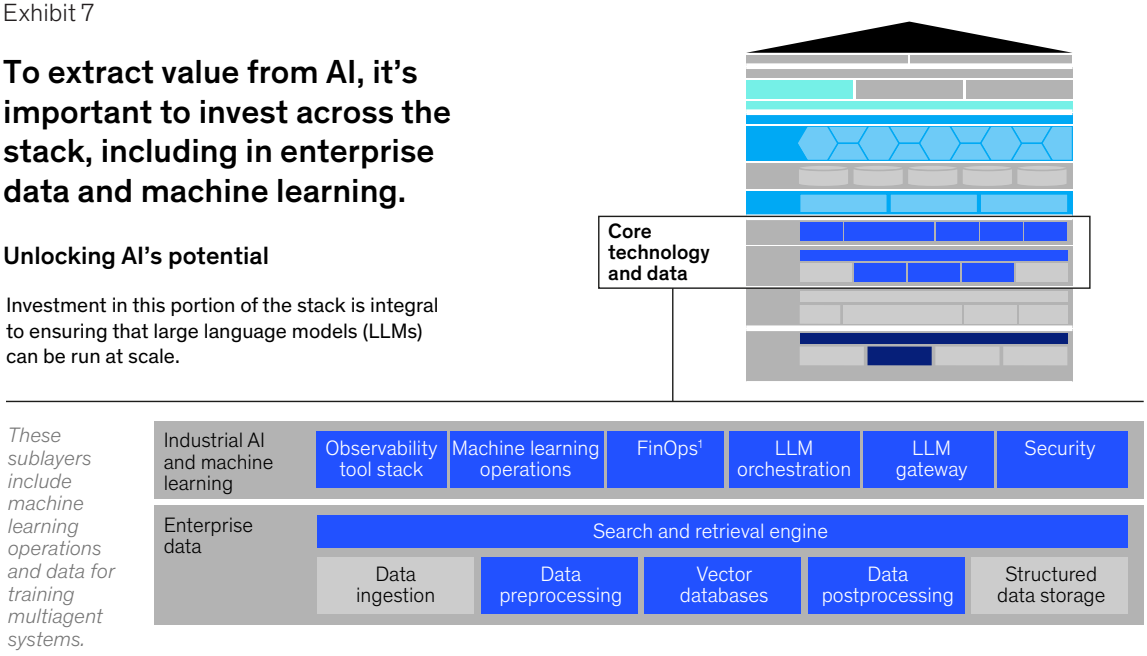
Investing in the foundations to enable AI value creation

Banks that unlock value from AI are making balanced investments across the entire AI capability stack. Envisioning this target-state AI stack is critical to ensuring that the right capabilities and innovations are built with an end goal in mind.

To build AI capabilities at scale, it's necessary to invest throughout the stack, including in the following crucial sublayers (Exhibit 7):

Exhibit 7

To extract value from AI, it's important to invest across the stack, including in enterprise data and machine learning.



[1]Financial operations, a framework for managing the operational costs of cloud computing.

McKinsey & Company

— *Industrial AI/machine learning sublayer.* This sublayer within the data and core tech layer provides the reusable tools and pipelines, equipped with a full suite of machine learning operations capabilities, needed to cost-effectively deploy and run LLMs at scale.

— *Enterprise data sublayer.* Another portion of the data and core tech layer, this sublayer is used to store, transform, and access large unstructured data sets, which are critical for training multiagent systems.

Sustaining and scaling value from AI

A successful AI transformation of a bank balances delivering a positive financial impact in the near term with building lasting AI capabilities for the enterprise.

After setting a business strategy with AI at its core and choosing the domains and subdomains to be transformed with AI, banks should focus on executing that transformation at scale, delivering value from the reusable components that can be created for one domain and then plugged into other domains as needed.

For example, a large bank is going through a multiyear transformation focused on improving performance and delivering analytics at scale with use cases including hyperpersonalization to target new customers and cross-sell to existing customers. The bank built reusable assets and an end-to-end analytics pipeline powering more than 50 machine learning models, developed a tool to consolidate customer leads from different sources and optimize them based on various factors, and built a fit-for-purpose, data-driven business operating model. Early results are promising, with projected revenue increases of 10 percent and usage of the resulting assets and framework in more than 150 use cases.

A transformation begins with one subdomain and the development of various use cases in that subdomain, moving through several phases, from minimal viable product to more sophisticated stages. As the transformation proceeds, reusable components from use cases in the first subdomain can be used in other subdomains (Exhibit 8). This process necessitates building and improving the AI stack in phases, as opposed to trying to create it all at once.

Setting up the technical elements outlined in the AI stack isn't enough to deliver value. Bank executives also have the following imperatives as they scale AI through the enterprise:

— *Setting up teams to create value, not just models.* To build capabilities for the future, banks should put in place cross-functional teams that seamlessly blend deep expertise in banking with technical expertise in technology platforms, data, and AI. Leaders should be thoughtful about creating transformation teams with adequate representation from all stakeholders and ensuring that everyone's incentives are aligned. One key consideration is embedding risk management expertise inside teams from the beginning and not treating it as an afterthought. Leaders should put emphasis not just on what is being built but also on how the new capabilities will be deployed and adopted by the bank's employees and customers.

— *Sustaining value.* Extracting value from AI is an enterprise-wide effort that requires centralized planning and coordination to deal with the inevitable bumps in the road that come with a transformation. To do this, many banks set up a central AI control tower and governance council that typically plays three roles:

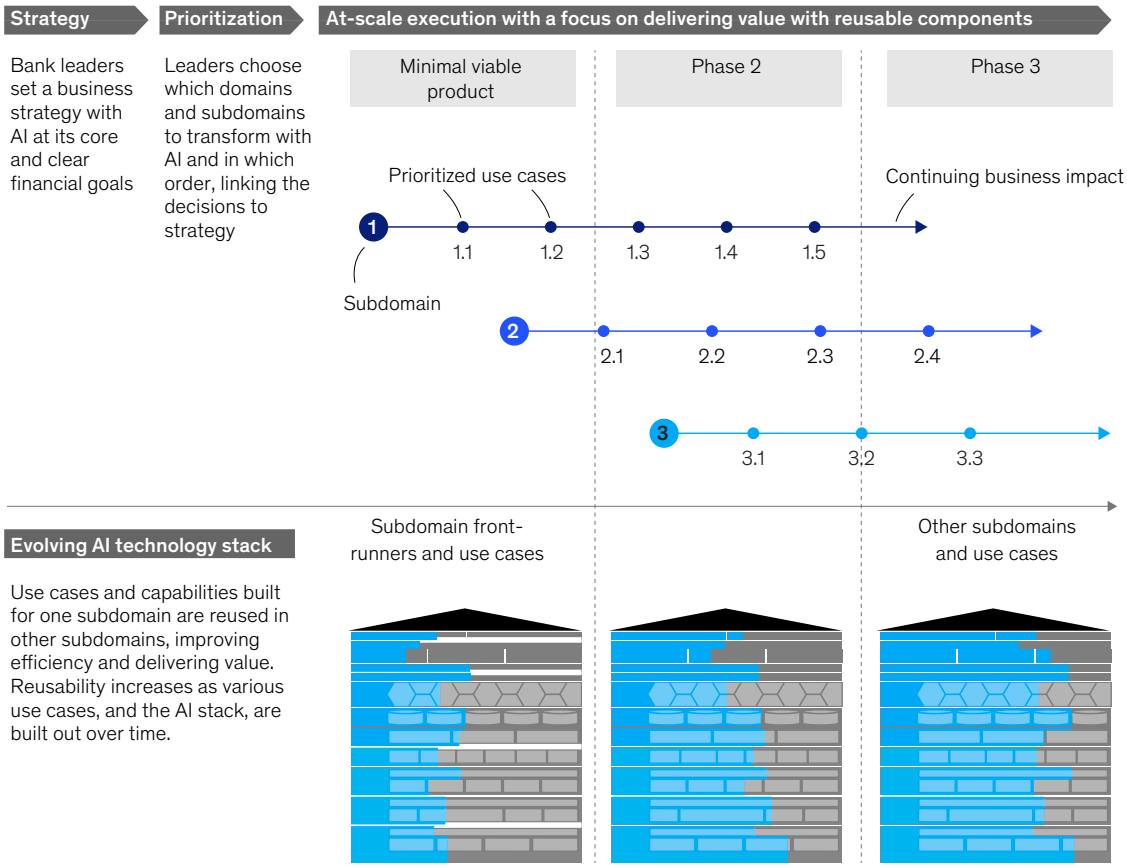
- *Custodian of the bankwide AI strategy.* In this role, the control tower tracks and monitors the value realized from various AI initiatives; frequently updates which areas to fund and which areas to stop funding, for instance, if experiments fail to yield results; and refines the bank's comprehensive AI road map.

- *Coordinator of key enterprise-wide decisions.* These include design choices about AI tools and the architecture stack,

Exhibit 8

In an AI transformation, reusable components from the first subdomain to be rewired can be implemented in other subdomains.

An example of an AI transformation of several subdomains at a bank



McKinsey & Company

such as what should be uniform and what can be federated. In this role, the control tower establishes a set of central standards and risk protocols to ensure cohesion across the bank's units.

- *Driver of enterprise AI assets' reusability.* Here, the control tower works with various business segments and functions, connecting dots across teams and accelerating the dissemination of best practices, and ensuring that AI capabilities built in one area of the bank, to the extent possible, are reused in other parts of the bank, saving incremental effort and accelerating time to market for product and service launches.

AI holds the potential to revolutionize the way business is done, but getting there will require more than mere experimenting. Organizations that employ the strategies outlined in this article can harness the power of AI to achieve scale and drive lasting, material value. Although only a few leading banks are currently generating material value from

AI transformations, it is possible that more could join them within the next few years.

Banks can use the following checklist to assess whether they are truly rewiring the enterprise to put AI first:

— setting up a comprehensive, bankwide vision for AI and measuring the ROI of AI investments (versus dabbling in AI within silos)

— taking a full-stack approach by blending gen AI with analytical AI and digital tools (versus seeking a transformation from gen AI alone)

— reimagining entire business domains, including optimizing internal processes and operations (versus deploying narrow use cases in isolation)

— using multiagent systems to automate complex workflows (versus training one model and expecting it to serve as a jack of all trades)

— ensuring the reusability of components (versus building each AI project from scratch with limited coordination)

Carlo Giovine is a partner in McKinsey's London office; **Larry Lerner** is a partner in the Washington, DC, office; **Renny Thomas** is a senior partner in the Mumbai office, where **Shwaitang Singh** is a partner; **Sudhakar Kakulavarapu** is a partner in the Bengaluru office; **Violet Chung** is a senior partner in the Hong Kong office; and **Yuvika Motwani** is an associate partner in the Gurugram office.

The authors wish to thank Antonio Castro, Archit Taluka, Rickard Ström, Romil Shah, and Vik Sohoni for their contributions to this article.

This article was edited by Jana Zabkova, a senior editor in the New York office.

Designed by McKinsey Global Publishing