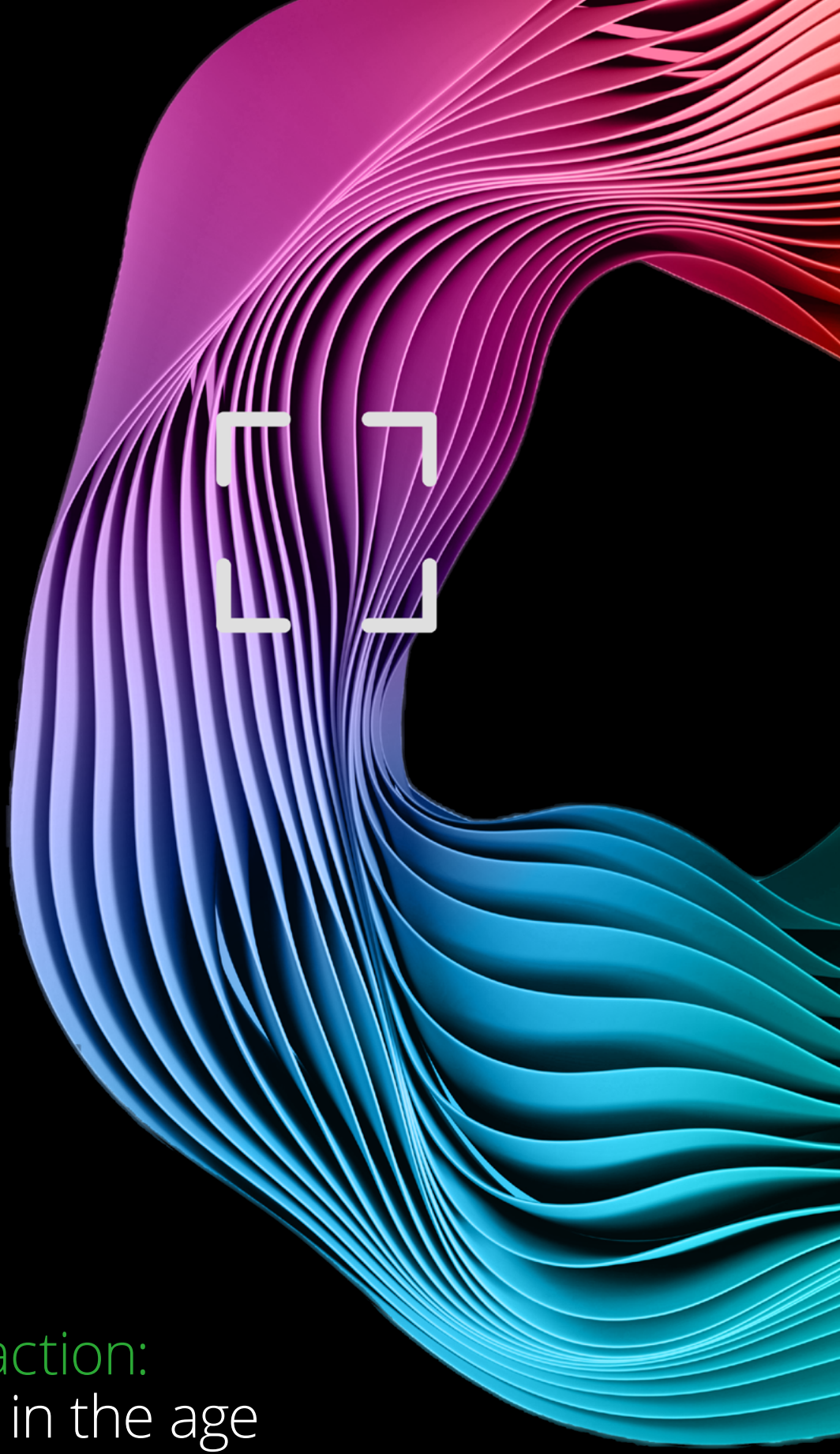
Deloitte.

Together makes progress



From trust to action:
Navigating risk in the age of agentic AI

Executive summary

The emergence of AI agents, which not only perform tasks but also identify, plan, and execute them with a higher degree of potential autonomy than familiar AI tools, calls for a renewed emphasis on established elements of the Trustworthy AI™ Framework. It also introduces new risks and challenges for organizations to confront.

- Agentic AI can function without constant human attention to internal decision points, but that doesn't mean it must. By building controls and visibility into the system, humans can retain confidence in what is happening and why.
- While agentic AI systems may ease some human knowledge burdens in specific subject matter, they will impose a new responsibility to amplify people's familiarity with AI so they can design, deploy, and use systems in trustworthy ways.
- A precursor to an agentic AI initiative is to assess the current state processes that it is going to manage: to confirm the quality of the data involved and to map the critical points at which monitoring and control will be consequential.
- Deloitte's Trustworthy AI Framework remains at the center of risk management, but incremental and refreshed risk management considerations are needed given the new risks that agents bring over stand alone AI or GenAI.

The rise of artificial intelligence (AI) agents and multi-agent systems—collectively known as agentic AI—has the potential to not only revolutionize organizations, but also redefine how humans and machines work together. Unlike the large language models (LLMs) or Generative AI (GenAI) tools many are familiar with, AI agents are reasoning engines that can understand context, plan workflows, connect to external tools and data, and execute actions to achieve a defined goal.[1] This makes agentic AI more an iterative technology than a net-new one, yet its power and implications call for renewed attention to trust, even from organizations that have already made strides in building trustworthiness into previous forms of AI.

A system that uses AI agents is called agentic AI, and a system that combines the actions of multiple AI agents into a larger process is multi-agentic AI. Both agentic and multi-agent AI systems not only generate outputs, but also make decisions about the tools they use and the processes they follow, with answers that may differ each time. To create trust, agentic AI has to do more than work reliably; it needs to be viewed as beneficial for users, as well as transparent and explainable in how it works.

Think of electricity: We all know that it exists and powers our daily lives but likely don't understand the ins and outs of currents and circuitry. But we trust electricity, because when we flip a light switch, it turns on without shocking us. New governance controls for agentic AI can similarly preserve the visibility and other qualities that make trustworthy AI possible. The first step organizations can take in conceiving those new controls is a fresh look at processes as they exist today.

More complex AI agents create more complicated risk and trust profiles

AI agents introduce more complexity into the human-technology trust relationship, as they exercise different levels of autonomy than other AI tools. A single AI tool like a chatbot may have the ability to transact, but an agent can also make cognitive reasoning decisions that more closely mimic human behavior. Instead of merely interacting with a user like a customer service chatbot does, AI agents are designed to reason and act on behalf of a user—a critical difference.

This notable distinction in agentic AI systems makes them more powerful than other AI tools, but also introduces increased risk and new governance considerations. Consider the leaps that happen from the use of single AI tools, such as LLMs, to the use of AI agents. Instead of automating tasks, the agent automates whole workflows, both creating and executing multistep plans to reach a defined end goal. These autonomous capabilities can transform how businesses and enterprises operate yet carry new risks—not because humans have been removed from the loop altogether, but because the relationship between people and machine processes inside that loop has changed. And as these operations grow more complex, the mechanisms to safeguard trust must evolve accordingly.

Trust in AI technology starts with trusting humans

To explore the trust implications of agentic AI, it's important to understand the foundations of trust: how humans form it and how we extend that trust to other people and, in turn, technology.

Trust is a human phenomenon, and the trustworthiness of non-human processes, such as AI, has to make sense in that context. Deloitte has identified four factors that inform people's sense of trust: **reliability**, **capability**, **transparency**, and **humanity**. As our authors Ashley Reichheld and Amelia Dunlop wrote in *Harvard Business Review*: "Think of it this way: If you go to a restaurant and find that it's unreliable (it lost your reservation), lacks capability (its food is poorly prepared), isn't transparent (it includes hidden surcharges on the check), and staff don't express humanity (they ignore your special requests), you won't trust it, and you won't return."[2]

If an institution like the restaurant can forfeit trust by falling short in these areas, so can an organization—or a technology. When that happens, trust is only the first element that's lost. Reputation or relationships can follow, or in the case of an organization, money. A study in *The Economist* found[3] that a company that loses trust can see its value erode by almost one-third in the short term.[4] On the other hand, companies that are leaders in trust can outperform others in market value by a factor of four.[5]

Note that because trust is a human quality, trust in AI must begin with humans who are conversant with how it works, experience it working consistently and reliably, see its relevance to their needs, and understand its implications. In Deloitte's recent State of Generative AI in the Enterprise survey,[6] 35% of the respondents said that the largest impediment to GenAI adoption over the next two years was the concern that mistakes caused by AI would lead to real-world consequences.[7] Additionally, 29% of respondents said that a general loss of trust due to bias, hallucinations, and inaccuracies could slow adoption rates. These concerns can be overcome if the technology's reliability, accuracy, and trustworthiness are improved; especially when implementing agentic AI in an organization.

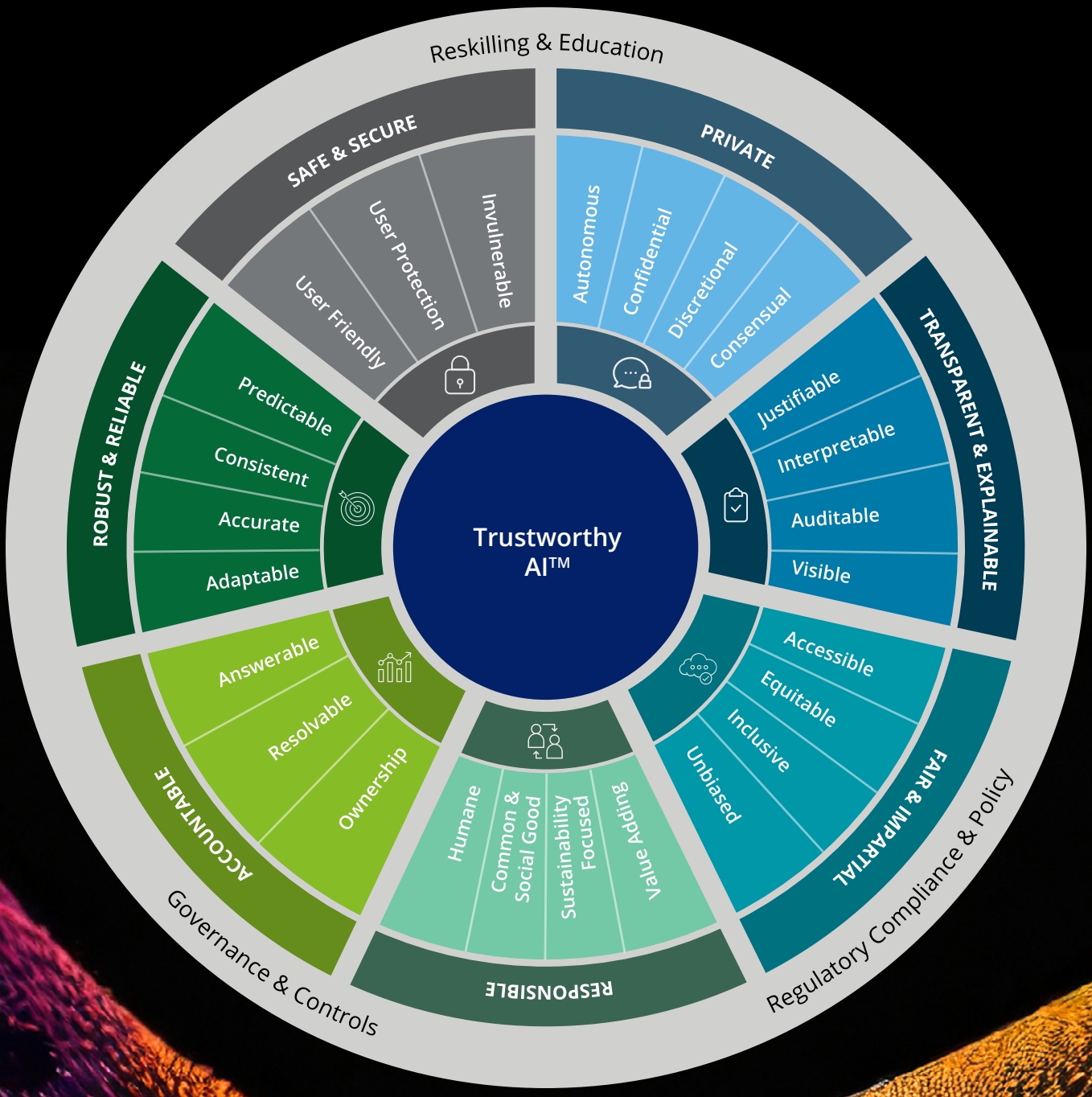
Using AI alone won't build trust—people need to trust both the technology and the organization that uses it.

A trust framework that applies across technologies

AI has evolved rapidly and iteratively, from machine learning (ML) to robotic process automation (RPA) to GenAI, and now agentic AI. At each step, machine processes took on tasks that required humans and institutions to trust their outputs. To make sense of that requirement, Deloitte developed the **Trustworthy AI™ Framework**.[8]

Each of the framework's seven dimensions apply to agentic AI just as they apply to other uses of AI. For example, personally identifiable information (PII), company data, and other sensitive information need to remain under control no matter what kind of information technology system handles them. Fair and impartial outcomes are vital across any instance of AI. However, there are other parts of the framework that take on new emphasis in the age of agentic AI. These include the need for systems to be **Transparent and Explainable**, **Accountable**, **Secure**, and **Reliable**, since agentic AI automates parts of processes previously under manual control. It takes close attention to the operations "inside the black box" for people and organizations to know what is happening (Transparent and Explainable), to be able to trace the movement of data (Secure) and decisions (Accountable) from step to step, and to have confidence in the outcomes of multistage AI operations (Reliable).

Why do some elements of the Trustworthy AI Framework figure more prominently than others in a discussion of agentic AI? The answers lie in the new approach: what it is and how it works.



Agentic AI amplifies familiar risks and introduces new ones both within and beyond existing trust frameworks
As we have noted, using established AI capabilities in new combinations the essence of agentic AI creates new trust issues and elevates familiar ones. Some of the new or heightened risks that can be specific to agentic AI include:

Runaway AI agents
Agents might perform malicious tasks or obfuscate steps to hide their footsteps

Data leakage and context amnesia
Memory corruption can lead to data leakage across users, or context amnesia might lead to insecure results

Misaligned learning
Systems might learn the wrong behaviors or use untrustworthy and unethical actions to achieve goals

Orchestration loops
Repetition or resource exhaustion can magnify errors

Context untraceability and forensics
Autonomous actions and nested permissions might lead to blurry accountability

"Confused deputy"
An agentic system with a high number of service identities, nested authorization, and/or inherited permissions can mask control

External dependency attacks
The external knowledgebase or external tools can be compromised

Agent supply chain
Agent components can be compromised, affecting the ability of agent security car metadata to share security insights, issues, and concerns

Of the seven dimensions that make up the Trustworthy AI Framework, here are some that merit fresh attention in the development and use of agentic AI.

Agentic AI and transparency

Transparency and explainability help users understand how technology factors into operational decision-making. In the work of AI agents, the outputs need to be explainable to the people that use them, and users need to understand how and why their data is used. Just as importantly, the people engaging with AI agents need to believe that these tools provide personal benefit. To confirm that AI drives the proper intent, organizations should design agentic AI systems beginning with the human need before the technology—taking a human-centric approach, prioritizing people's desires alongside business requirements, and evolving the technology from there. Ultimately, an agentic AI system should be transparent in an end-to-end evaluation that illustrates the logic of its planning and routing decisions.

Agentic AI and accountability

When an AI agent or multi-agent system is at work, their actions and decisions may not be visible in the moment. That makes it important for organizations to maintain a record of those elements for use in later reviews, with a central trust framework as the standard. What sets an agentic AI model apart from a single AI model is the existence of routers: components of an AI agent that direct the model from step to step. It's essential to understand and test the routes an agent chooses—and why—since the agent is autonomously working through a decision tree. Trusting the way it does so requires someone on the human side, either individually or institutionally, to own responsibility for the way the agent performs and to be accountable for the outputs.

Without that understanding, the structure of agentic AI amplifies its potential to jump the guardrails of accountability. Unlike a single AI operation, it can perform unintended tasks or generate outputs without clarity as to what informed them. Accountability is not a new challenge in AI, but it applies to agentic AI with greater intensity.

Agentic AI and security

AI models can catalog potential risks and follow rules to avoid them, but it takes humans to anticipate and understand those risks. Humans define the potential risks, humans design the tools, and humans remain in the loop as those tools operate. Their role in keeping agentic AI secure is different from what came before, but it is just as important. Failing to recognize that, and treating agentic AI as a fully autonomous architecture, is where security lapses can cause unauthorized access and unwanted manipulation, resulting in a loss of trust.

Agentic AI and reliability

People trust machine outputs the same way they trust human action: gradually and based on concrete experience. Systems will begin by promising trust, but in the end, they need to promote trust through consistency in their outputs. When that happens, humans will keep using them, and trust can continue to grow if results remain acceptable. For this reason, adopters of agentic AI may opt to begin with low-stakes functions and outputs, experience the results for a time, then expand the technology to address higher-stakes decisions and actions.

Where human and machine trust come together in practice

Knowing the importance of exploring these issues, Deloitte built an AI assistant for our professionals' internal use. In our case, we also used it as a test case for how to build trust in technology among our workforce. The assistant in question used GenAI, not agentic AI, but as we have seen, the two approaches present similar trust challenges—though with agentic AI, they can be more acute because of reduced moment-to-moment human control. Tracking users as they progressed from low trust at the outset to a more confident view some months later, we identified several steps that have the potential to enhance AI trust in other organizations.

- **AI superuser profiles** to show how people used the tool and the ways they've benefited
- **Public Q&A sessions** where people can ask the technology team about how the tool works and how it uses data
- **"Prompt-a-thons"** in which people can hone their ability to write effective queries, or "prompts," for AI tools to act on
- **Community forums** to share trust insights and system updates as people gained experience using AI

Perhaps the biggest lessons we learned were to lay the foundations of trust before launching a new AI technology, to build in design elements that promote the factors of trust, and to devote resources to formal AI training. In our Deloitte pilot, the tool in question was GenAI, yet these principles have the potential to prepare our workforce and shape a better future experience with agentic AI in much the same way.

What to do now: Trust considerations for agentic AI

As it emerges and matures, agentic AI will present questions of accuracy, and real-life experience will begin to answer them, just as with earlier AI variants. For each use case, designers should perform a risk assessment, define the logic that can address the identified risks, and document the ways the system applies the resulting governance in its operation. This amplifies the need for monitoring and auditing processes and outputs.

Consider your organization's risk tolerance.

- Defining a risk tolerance for agentic AI starts with asking critical questions and devoting resources to finding the answers in real time. For example, an AI agent can identify the need for a targeted communication, devise the appropriate message, and send it to a human recipient. But will an organization trust that agent from day one to send emails to customers or suppliers? Should the agent start by generating emails only for internal audiences? Or should an early implementation generate the email for a human to approve before sending?

Establish standards for process readiness.

- Which organizational processes are mature enough that you can begin to build agents to take them on? Which ones need work?
- Review your organization's current data sources, accuracy, consistency, structure, and availability. Industry observers characterize the general state of data in large enterprises to be average at best. Many organizations may feel they've already invested in a data transformation for AI, but whatever it takes to achieve, agentic AI may raise the standard again.
- Human processes are systems worth evaluating too. Where are the decision points, hinges, and trust gaps? How can the design of an agentic AI system, including monitoring, acknowledge and address them? Waiting to address these questions until a system is up and running may limit the ability to promote trust.

See the big picture first.

- Set rules for human oversight that correspond to the AI agent's new role. For example, if an AI agent is automating many of the tasks a junior financial analyst might perform, can the junior analyst's job evolve to include monitoring and analyzing the agent's outputs?
- Be realistic and transparent in what you expect from agentic AI and the ways you use it. The fastest way to erode trust is to set an expectation and fail to meet it.
- Consider an end-to-end evaluation of your enterprise AI policy to help build trust by confirming that AI systems are designed, deployed, and operated in a manner that aligns with ethical standards, regulatory requirements, and organizational goals.

A renewed focus on trust and understanding

Agentic AI has the potential to redefine the relationships between people and technology. AI agents can amass business or subject knowledge quite rapidly, based on how humans train the models. But to manage that process with trust, another form of human expertise comes to the fore: Instead of, or in addition to, academic or business knowledge, people will need the ability to understand AI systems and the ways they work.

Ultimately, trust is an asset that moves among people and institutions. If your organization reaches an internal level of trust with agentic AI, it still may face the challenge of winning over external stakeholders. It may not be visible outside the organization that agentic AI is in use; some parties won't see the distinction between that and, for example, machine learning of GenAI. Others may interact with an organization unaware that AI is even in use. But the system's outputs are the organization's outputs—and they shape its reputation for trust no matter how they were created.

Technologically, agentic AI is more an iterative evolution than a transformation revolution, but its application will almost certainly revolutionize how work is done in the future. Inside your organization, you and your stakeholders will see and feel the change. This is ultimately a realignment of existing AI tools into new process structures. That means it will require a renewed application of trust awareness and safeguards that you already understand. A careful mapping of the systems agentic AI will run can help inform a reexamination of the critical points where monitoring and control are important. From that starting point, organizations can look with confidence toward a new era of advanced automation.

---

1. Vivek Kulkarni et al., "Prompting for action: How agents are reshaping the future of work," Deloitte, November 2024.
2. Ashley Reichheld and Amelia Dunlop, "4 questions to measure—and boost—customer trust," *Harvard Business Review*, November 1, 2022.
3. *The Economist*, "Getting a handle on a scandal," March 28, 2018.
4. Reichheld et al., "4 questions to measure—and boost—customer trust."
5. Ibid.
6. Jim Rowan et al., *State of Generative AI in the Enterprise: Quarter 4 report*, Deloitte, January 2025.
7. Reichheld et al., "4 questions to measure—and boost—customer trust."
8. Deloitte, Trustworthy AI™, accessed June 2025.

Deloitte.