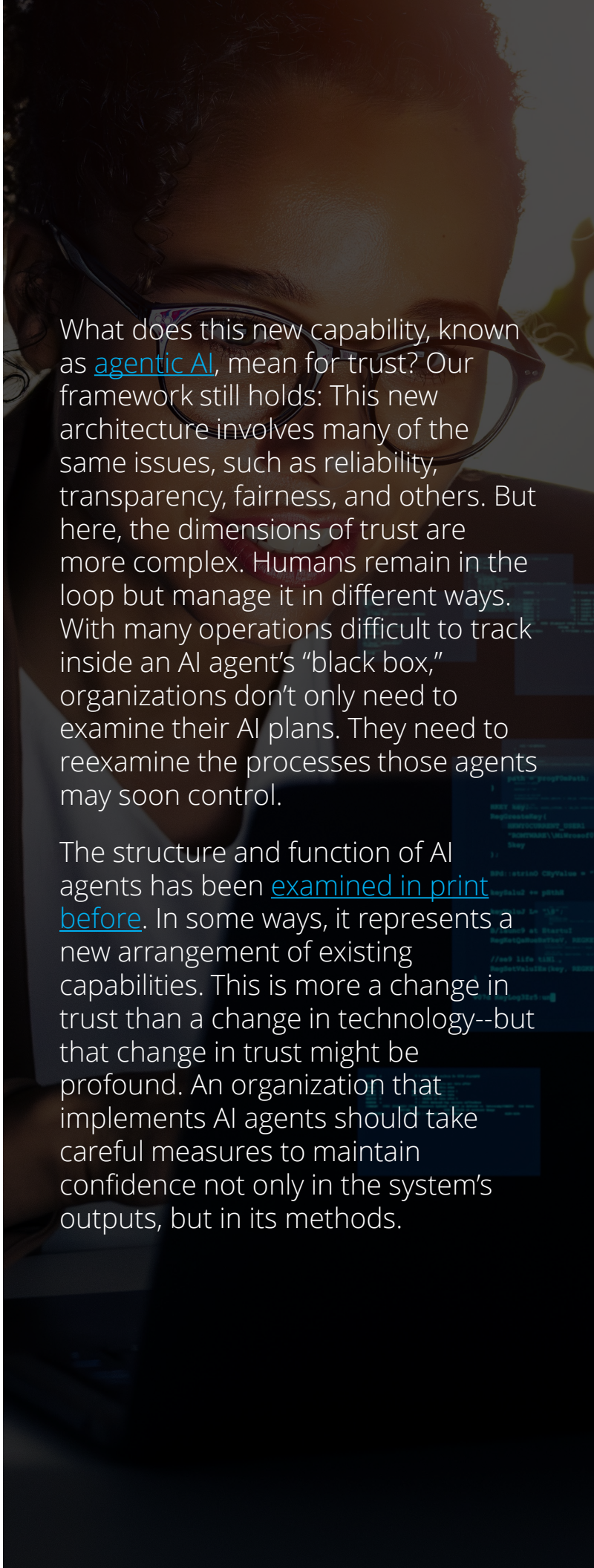
Deloitte.

Risk and Trust in the Age of Agentic AI

Risk and Trust in the Age of Agentic AI

Trust has been a concern ever since artificial intelligence (AI) first operated. Deloitte established the Trustworthy AI Framework to identify and address the many aspects of that challenge, and the expanded capabilities of Generative AI placed a new emphasis on it.

Now, another evolution is prompting a fresh look at trust: AI agents are reasoning engines that can plan workflows, connect to external tools and data, and execute actions to achieve a defined goal. Instead of merely interacting with the user, they are designed to reason and act on behalf of the user.



What does this new capability, known as agentic AI, mean for trust? Our framework still holds: This new architecture involves many of the same issues, such as reliability, transparency, fairness, and others. But here, the dimensions of trust are more complex. Humans remain in the loop but manage it in different ways. With many operations difficult to track inside an AI agent's "black box," organizations don't only need to examine their AI plans. They need to reexamine the processes those agents may soon control.

The structure and function of AI agents has been examined in print before. In some ways, it represents a new arrangement of existing capabilities. This is more a change in trust than a change in technology--but that change in trust might be profound. An organization that implements AI agents should take careful measures to maintain confidence not only in the system's outputs, but in its methods.

Familiar AI risks, only larger and more numerous

Instead of "only" generating answers based on human-fed prompts, agentic AI makes decisions and directs other processes autonomously. That puts humans in a different relationship with the step-by-step function of the system. Trust issues an organization may already have considered in the use of machine learning or Generative AI may reappear, in a familiar form but at a larger scale.

AI agents may never be fully autonomous, but the degree of autonomy they do represent means organizations likely need a fresh take on the human-machine relationship.

Points to consider:

1. If the output of one internal process becomes the input for another, it can be difficult to identify where errors or hallucinations originated. An error early in a sequence may become magnified by the end.

2. The learning and continuous adjustment inherent to AI makes the system non-deterministic. Each operation and its output is different from the last. That's one of AI's many valuable attributes. But risks multiply when it becomes harder to answer the question: Different *why*?

3. Governance doesn't only control processes; it is a process. AI can make it more effective. But an insufficient approach to trust and control has the potential to amplify error—or conceal it.

4. When humans have a smaller role in the loop, they have less need for some of the AI skills they have honed in recent years. But they will need new ones—not only to manage agentic AI, but also to take on the higher-level strategic jobs the new architecture frees them to address.

The basis for trust is there

Humans built the processes you use today, and established the protocols that make them reliable. That means humans know where the critical points and guardrails are—and when it's time to implement agentic AI for those processes, it will be humans who design those systems and know where to focus the required controls.

Points to consider:

Familiar principles of trust do apply to agentic AI, but the work of data management and controls design may take "greenfield" efforts—applying data cleansing and transformation to the new architecture and to categories of data that previous AI efforts did not address.

Before, AI used data to perform actions. Agentic AI uses data to plan and decide as well. That means new categories of data may need cleansing and transformation that didn't happen before.

The feedback loops that AI agents use to learn and refine their operations are a key focus of trust. Agentic AI architects need to anticipate and test all possible "routes" of action within a system.

Right now

The risks that may imperil trust in a future agentic AI system are likely to concentrate in the critical points of the processes as they exist today. For any operation your organization may assign to AI agents, catalog the hinges and decision points where control and monitoring will make the greater difference.

Points to consider:

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 begin the work of establishing governance and controls for later agentic AI design and deployment, consider three broad steps:

Assessment: Inventory applicable use cases and determine the risks of each one.

Mitigation: Apply risk logic and devise controls and guardrails for each use case, including the use of simulations.

Monitoring: Continually track the model's performance and retrain it as needed. Don't only flag anomalies; optimize processes wherever possible.

Standing on the shoulders of past understanding

Because of the way it works, agentic AI might not capture the public imagination the way Generative AI has done. But 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 re-examination of the critical points where monitoring and control are most important. From that starting point, organizations can look with confidence toward a new era of advanced automation.

Look out for our upcoming white paper on agentic AI's trust and risk implications, with practical examples and use cases for organizations to consider.

Contacts

Beena Ammanath
Managing Director
Global Deloitte AI Institute Leader
Deloitte Consulting LLP
bammanath@deloitte.com

Acknowledgments

Amandeep Singh, Ashley Reichheld, Derek Snaidauf, Jim Rowan, Mike Crowthers, Prakul Sharma, Scott Holcomb

Deloitte.