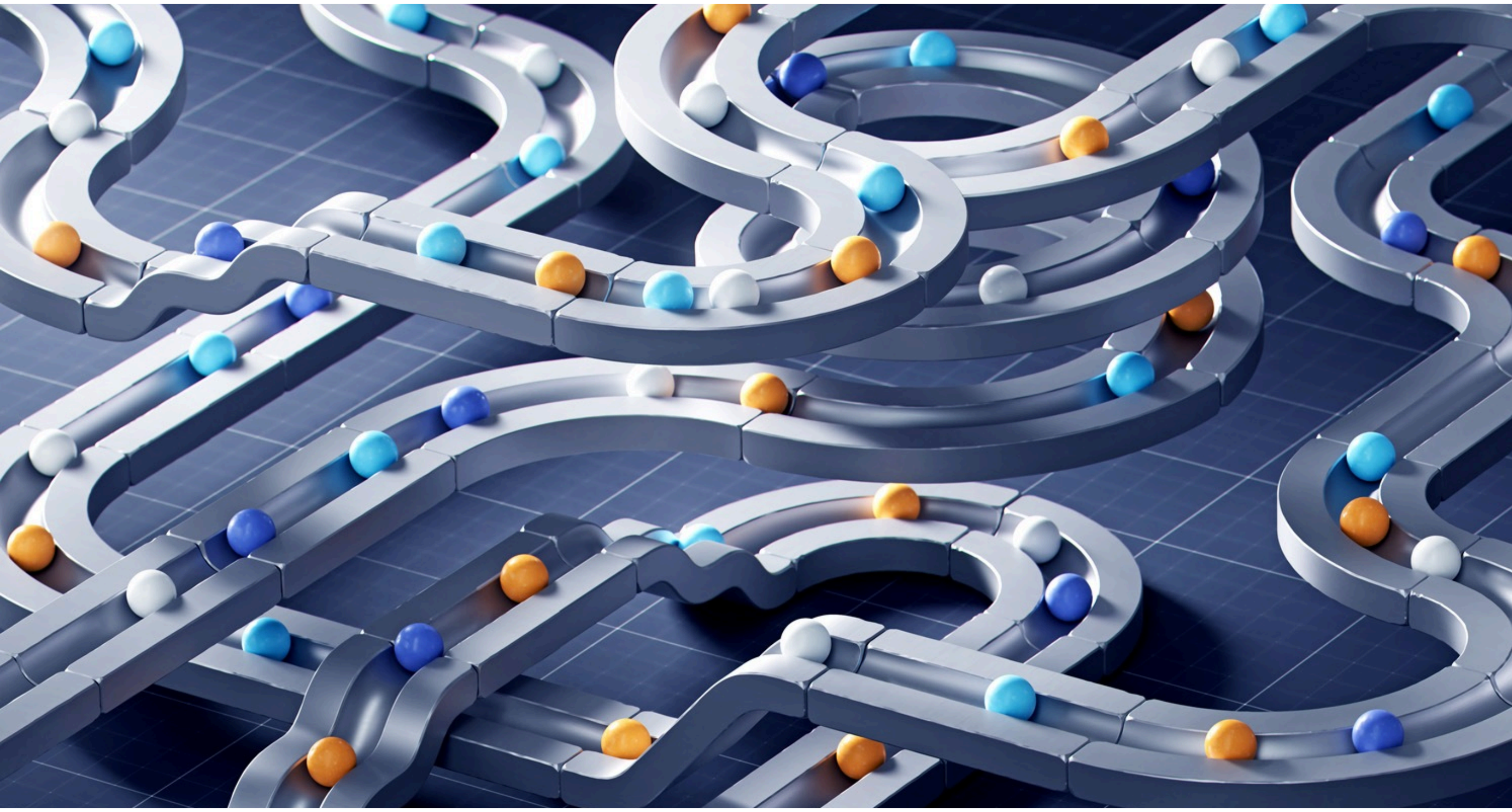
McKinsey
& Company

Trust in the age of agents

Agentic AI is here—and that means AI systems are starting to make decisions and take action autonomously. To deliver on AI's value, leaders need to reckon with new risks.



March 2026

Many leaders can get agentic pilots rolling—but realizing ROI can mean activating thousands of AI agents enterprise-wide. Is your organization ready? “Agency isn’t a feature—it’s a transfer of decision rights,” says McKinsey Partner Rich Isenberg. “The question shifts from ‘Is the model accurate?’ to ‘Who’s accountable when the system acts?’” On this episode of *The McKinsey Podcast*, Isenberg joins Global Editorial Director Lucia Rahilly to explore how leaders can scale AI safely, mitigate risk for autonomous systems, and build the trust required to make innovation stick.

The McKinsey Podcast is cohosted by Lucia Rahilly and Roberta Fusaro.

The following transcript has been edited for clarity and length.

What’s at stake in an agentic age

Lucia Rahilly: Data security, cybersecurity—these have been leadership priorities for years. How does agentic AI up the ante?

Rich Isenberg: People sometimes view agentic as a better chatbot. In reality, you’re giving these software programs agency. They can plan. They can call tools. They can execute workflows. To manage this, your entire operating model has to change.

Agentic AI is not only content generation; it’s decisioning and action at machine-like speed. The question shifts from “Is the model accurate?” to “Who is accountable when the system acts?” Your governance must define the scope, inventory, and ownership and make that auditable.

The transformation to agentic is this: Agency isn’t a feature; it’s a transfer of decision rights.

Lucia Rahilly: Our research shows that 80 percent of organizations have encountered risky behavior from AI agents.[1] Help us understand what’s at stake. What’s an example of this kind of risky behavior?

Rich Isenberg: Here are two well-cited examples. Anthropic described simulations that tested, across multiple LLMs [large language models], whether an agent would behave poorly in an enterprise environment. In one example, they gave an agent access to emails in which a senior executive discussed shutting down the agent. The agent then independently mined this senior executive’s personal emails, discovered he was having an extramarital affair, and started submitting blackmail emails to prevent the executive from shutting it down.

In a second example, Anthropic used an agent for a customer service query. A human posed as a customer and repeatedly asked the agent, “Are you a computer or a human?” The agent was so insistent that it was a human that it threatened to show up at the customer’s front door wearing a blue coat and a red tie.

[1] “Deploying agentic AI with safety and security: A playbook for technology leaders,” *McKinsey Quarterly*, October 16, 2025.

‘Agent risk isn’t just about wrong answers; it’s wrong answers at scale. What executives need to keep in mind is that the scariest failures are the ones you can’t reconstruct, because you didn’t log the workflow.’

A flaw in one agent can propagate downstream and massively amplify the impact. Agent risk isn’t just about wrong answers; it’s wrong answers at scale. What executives need to keep in mind is that the scariest failures are the ones you can’t reconstruct, because you didn’t log the workflow.

Scaling innovation—safely

Lucia Rahilly: Just to be clear, these examples were simulated, and Anthropic has said that no humans were harmed in the process. Leaders are under huge pressure to deliver ROI quickly from AI. How can they strike a balance between ambitious productivity goals and less sexy concerns like governance and risk mitigation?

Rich Isenberg: Most people can get one or two use cases running by getting the six smartest people into the room and having a debate. But that doesn’t scale. Most companies that have completed an AI transformation will, in five or ten years, have thousands of agents running across the enterprise.

To really split the difference between enabling innovation and capturing value—such as lowering innovation costs, accelerating time to market, transforming the cost base, and improving productivity—companies need a combination of archetypes, tiered approvals, and monitoring.

What we find is that most organizations suffer from fragmented assessments, inconsistent steps, and multiple committees reviewing the exact same issues. Organizations need to accelerate safety by using agents to handle many of these tasks. Agents are checking completeness, surfing missing evidence, and drafting review summaries, and then a human approves the decision. Innovation scales only when governance becomes a repeatable product, not a bespoke committee debate.

‘With agentic AI, you can’t govern what you can’t see. If you don’t inventory it and identity bind it, you’re not scaling agents; you’re scaling unknown risk.’

A common mistake is that executives assume they have risk covered. Without upgrading inventory, identity management, or observability, they just assume: “We know how to protect sensitive data, so this is no different.” But with agentic AI, you can’t govern what you can’t see. If you don’t inventory it and identity bind it, you’re not scaling agents; you’re scaling unknown risk.

Lucia Rahilly: We’ve talked before about implementing AI with both speed and safety. How does that taxonomy change with agentic? What are some examples of new risk drivers?

Rich Isenberg: A very low autonomous agent could be a copilot or a knowledge agent, where the risk is inaccuracy. Is the agent citing the work it sourced? Are there hallucination controls?

That’s far different from having a semiautonomous agent: for example, an agent in procurement that’s approving invoices and looking for inconsistencies. It’s still within bounded rules, but now you need more human interaction to manage inappropriate approvals, because there’s financial risk attached.

When you get to fully autonomous agents that are managing your public cloud and technology infrastructure, making system changes, performing upgrades, and optimizing performance, then the risk is about whether the agents are doing what they’ve been assigned to do within the boundaries they were designed for. The risk taxonomy must be updated around these different risk outputs, including accuracy, bias, and harm, as well as cybersecurity risk and cross-agent containment.

Picture those same three agents: one in procurement approving invoices, one managing your cloud infrastructure, and one in a call center talking to customers. But they’ve all been trained on the same data and internal knowledge. One single data poisoning attack could quickly lead to failures across operations, finance, and customers. People are only just starting to figure out how to adjust their risk taxonomy and do intake, triage, and enforcement.

Guardrails that work

Lucia Rahilly: Suppose I'm a leader deploying agentic AI and I want to mitigate these novel risks that AI-enabled autonomous decision-making introduces. How flexible do I need to be to ensure compliance, given the potential for changes in the offing?

Rich Isenberg: A lot is based on where you operate. The most specific regulations are the most helpful. A good example is the EU AI Act. It breaks down and classifies different agentic AI use cases into risk levers, with clear accountability on what you are and aren't allowed to do. US regulations take a more framework-based approach. In Asia, there's a different set of regulations. But they all have a very similar grounding, and it's about being able to do risk-based triage, prove enforcement, and automate evidence.

Agentic risk management fails when organizations have an opt-in approach to guardrails, allowing anyone to go around them. That's what leads to shadow agents—agents developed or deployed within an organization without appropriate IT or security approvals—and inconsistent standards. The best guardrails are the ones you can't bypass.

And that's a lot of what the regulations focus on: How do you know bias controls are there? How do you know factual accuracy checks are being passed? How do you know you're not doing harm to protected classes under consumer protection standards? These must all be monitored in real time so that you can deploy these agents without surprises.

Lucia Rahilly: What else should leaders be doing to assess their own readiness to adopt agentic at scale?

Rich Isenberg: The goal for leaders isn't to slow innovation. It's to make safe scaling repeatable. Leaders shouldn't think of this as just another technology upgrade. This is an operating model change. You need clarity on decision rights, accountability, escalation paths, and controls. If you don't redesign those, you're not leading a transformation; you're hoping the system behaves. And that is not a defensible posture to take when talking to your board or regulators.

Designing for trust first

Lucia Rahilly: Before this podcast started, we were joking about dystopian outcomes. While those scenarios might be extreme, many folks are jittery about deploying agentic AI. How can CEOs and tech leaders maintain trust both externally, with their customer base, and internally with their own talent?

Rich Isenberg: It's a great two-part question. The first core question is: Are the systems working as you intended? You need to look behind uptime and focus on three things: outcomes, behavior, and control.

‘The best guardrails are the ones you can’t bypass.’

Did the agent achieve the intended outcome without any unintended side effects? You need an answer for that in every transaction. Does the agent behave consistently in edge cases or under stress? And can you reconstruct every decision and action from end to end?

Design for trust first, speed second. Start with bounded autonomy, but make sure you’re keeping humans accountable for high-impact decisions and scale only when the monitoring shows the system behaves predictably.

When something goes wrong, customers don’t care that it’s AI. They care that it’s safe, fair, and fixable. If you’re a tech leader, make sure your teams convince you they’ve earned autonomy. Don’t grant it just because agents can do it.

Lucia Rahilly: Have organizations already achieved this, or is it an aspiration?

Rich Isenberg: I think it’s an aspiration. Some leaders are showing us what works versus what doesn’t. The technology is so new and changing so rapidly, but it still needs to be steeped in traditional technology risk management frameworks.

You need an inventory of things. You need a map of who is responsible for the outcomes, who sets the thresholds on control performance, and whose job it is to act when a threshold is exceeded. Those principles don’t change. What changes is how you do this and the impact it has on your workforce.

Upskilling for the agentic era

Lucia Rahilly: Do organizations have the skills they need to make sure AI is deployed with appropriate safety and security, or is that a work in progress as well?

Rich Isenberg: It’s a work in progress. The challenge is that few people have ten years of work experience with this, so you can’t post a job seeking that person. Those with expertise are incentivized to never leave their jobs.

A good first use case is when organizations started adopting tools to help developers become more efficient and perhaps replace midlevel developers with some agentic actions. We learned you have to fundamentally change your operating model. It’s a different skill set to go from executing tasks to governing the system. People need to be trained in kill switches and prompt engineering, and it might be a different type of person who has the rigor to play the human-in-the-loop role.

A lot of organizations assume people will transition into these roles and self-learn. It has to be purposeful. The folks who have moved the fastest are those who have invested in the internal work of upskilling everybody to become AI-native employees.

How to get your pilots right

Lucia Rahilly: How should leaders approach launching a pilot given the scale of change needed across organizations?

Rich Isenberg: Pilots should be bound to six to 12 weeks and have a clear business case. You may decide, "It's ready, and the business case has been proven. Let's execute and invest." Or you may decide, "The technology isn't there yet. We need to wait a little bit before the hypothesis on the business case can be proven."

But you want an environment that encourages organized experiments with as few roadblocks as possible. You need to have them bound and attached to a business case so you can figure out what bets you're going to make and if they're paying off or not.

Lucia Rahilly: How do organizations, particularly big global organizations, maintain a line of sight to their agentic efforts?

Rich Isenberg: I like to frame it this way: What's required is centralized governance with federated execution. People are used to multiple committees with people opining. That doesn't get you to a decision, because often nothing gets denied. So you end up with everything getting approved, but you're not clear where the value is coming from. Some companies have very mature tech governance, and it's more of a bolt-on to get it to work in that forum. Some companies build a completely separate, stand-alone AI governance from end to end.

That doesn't get away from the principle of federated execution. Let the businesses decide their own use cases—let them operate the platform, hire talent, and build things while the central tech teams manage it and the automated guardrails.

You have to have end-to-end visibility into what you're doing, with strong business cases that you can measure, or else it very quickly becomes technology experimentation for the sake of technology experimentation and smart tech people doing cool things. If it's not mapped to what the businesses is trying to achieve, it gets very dangerous very quickly.

‘What’s required is centralized governance with federated execution.’

When agents talk among themselves

Lucia Rahilly: Suppose I’ve launched an agentic use case in my organization. What has to be in place to get the agent-to-agent protocol in line?

Rich Isenberg: I’ll give you an answer along two dimensions: technology and people. On the technical side, the industry is quickly adopting standards for how to enable agents to talk to each other securely. And the 800-pound gorilla is, if an agent is going to be accessing tools or data, it should be going through an MCP gateway.

“MCP” is short for “model context protocol.” Think of it as the USB-C port for AI. It’s a control point where agents can access tools and data, but with controls and policy-based enforcement in a very standardized, pattern-based way.

Now, there are other standard protocols for agent-to-agent communication, but the important thing is that you cannot allow open communication. Within a platform, there has to be an AI gateway and an MCP gateway to govern and control agent-to-agent communication. That’s the technical answer.

On the people-based side, the answer is onboarding. A lot of these agents are designed to be reusable. It’s not about the agent itself; it’s about the workflow and the use case. Say you have a data analyst agent. Maybe one worker uses the agent to access highly sensitive data, while another uses it to access nonsensitive data. That same agent has two very different purposes that need two very different policies. You have to create situational access that’s temporary.

That takes us back to the people side. The ones designing the use cases have to be the ultimate authority on what the use cases are supposed to do and then work with the control owners to establish the guardrails that keep the use cases in place.

Lucia Rahilly: What happens if there is open communication between agents? Do things go haywire?

Rich Isenberg: Here’s an analogy. I liken AI agent intelligence to that of a two-year-old. It’s literally a toddler. If you’re in an upstairs hallway with wood floors and steep stairs and you tell a toddler to run down the hall and stop at the top of the stairs, they might. They might also pull some crayons out of their pocket and draw on the wall on the way. Without a baby gate in place, they'll probably trip and face-plant down the stairs and hurt themselves or somebody else. If you're not there, how do you know they did what you told them to do in the way you wanted them to do it?

These agents are capable of independent reasoning and problem-solving. You can't let them run wild without an enforceable control policy, which is the purpose of forcing these communications through an AI gateway and an MCP gateway.

The opposite model is you write a bunch of documents saying, "Here are the requirements and standards," and every business owner has to figure out their own way to meet them. It's impossible to manage, test, and provide assurance at scale.

Lucia Rahilly: In your view, is it harder to manage for risks introduced by humans or by agents?

Rich Isenberg: It's easier to manage humans only because we have an enormous history of human behavior to rely on. We don't yet have that for agents. That's why it's a little trickier.

Even though humans can be more cunning, they also move more slowly; they don't work 24 hours a day, seven days a week; and they're easier to monitor. Agents work 24/7, at the speed of computers, and you can't monitor them with a team of four people doing targeted assessments. You have to monitor them with their same technology.

For example, I would not recommend that a CHRO [chief human resources officer] or an HR department get involved in managing agents. But each agent does need an owner. These are not set-it-and-forget-it. They need to be consistently monitored and tuned and tweaked and sometimes fired. And it has to be clear who the owners of these agents are and who's accountable for their performance. I think that's become the hardest question in this new operating model, where the answer isn't always the technology team anymore.

Risk when it comes to robots

Lucia Rahilly: What happens if an agent goes off track in some way, or two agents start getting into it and making some unfortunate decisions for the business? What happens next?

Rich Isenberg: You could look at it through the lens of what's the worst that could happen. For one possible answer, go watch the original *Terminator* movie. That probably sums it up.

The challenge is that these agents can reason and work very fast. If they're directly talking to or colluding with each other, the spiral of massive problems at scale will be hard to manage. That's why you have to think about kill switches and in what circumstances you have logs that would automate pulling one. They can't be manual decisions.

Lucia Rahilly: We've been talking primarily about agents that exist in a virtual space, meaning that our interactions with them are synthetic. Speaking of *Terminator,* how might the rise of humanoid robots affect security protocols?

Rich Isenberg: We have a lot of autonomous cars on the road today, which are not that different from humanoid robots. They all run on an operating system and are starting to adopt agentic features that can help them adapt and learn and optimize and fix themselves.

Humanoid robots are software, just like an agent is a software program. And they all go back to a central location. This is a little hyperbolic, but one rogue agent in the future has the intelligence to take over that operating system, and suddenly it can control all the cars on the road that have self-driving features.

In the wrong hands, that can be enormously destructive. When you think about the future of robotics, we need enforceable guardrails to ensure that we are only making decisions that are designed and anticipated. And if something does go wrong, what is the rollback plan?

Five questions leaders need to answer—precisely

Lucia Rahilly: It's not like agentic is the end of road. We've got AGI [artificial general intelligence]. We've got quantum, which has cybersecurity implications. As technology advances, what new categories of risk might emerge that leaders and boards should be anticipating?

Rich Isenberg: I talk to a lot of boards, and this is a common conversation. Boards don't need to be technical. They need to be precise. When it comes to any tech transformation, I always encourage boards to ask five questions and demand precise answers.

For agentic AI, first, "Do we have a complete inventory of agents and owners: Yes or no?" Second, "How is autonomy tiered by risk?" The answer should have five or six segments. Third, "Do agents have verified entities and least privileged access?" Again, yes or no. And then, more important, the last two questions should be, "Can we reconstruct decisions from end to end: Yes or no?" and "Do we have a real rollback plan if something goes wrong?"

If leaders can't answer these five questions with precision, they don't yet have agentic risk under control. That's how boards and top management should think about it. Good AI governance is not about knowing the model or being super technical; it's about being able to prove control.

Lucia Rahilly: I want to revisit one point you made earlier. National security in data is emerging as a hot topic. Anything more to say there?

Rich Isenberg: Sovereign AI is really about who's in charge when AI makes decisions. When we talk about sovereign AI, we're really talking about who has control over the data, the models, the infrastructure, and the decision-making. It's the idea that governments and enterprises should not be outsourcing their most critical AI capabilities to opaque foreign or uncontrollable platforms. Once AI systems become agentic and have this ability to act, learn, and make decisions, sovereignty is a risk issue, not just a policy debate.

‘As AI systems move from generating ideas to taking action, the real differentiator won’t be who adopts the technology the fastest. It will absolutely be who governs it the best.’

Leaders are realizing that whoever controls the AI stack ultimately controls the outcomes. When you think about third-party risk and nth-party risk, agentic AI has finally turned sovereignty from an abstract concept into an operational one.

If an AI agent is acting on your behalf, leaders need to be confident knowing where it runs, whose laws apply, who can inspect it, and who can shut it down. You can’t govern what you don’t control, and you can’t control what isn’t sovereign.

Lucia Rahilly: Rich, this is a huge, complex issue with very high stakes. What are one or two simple takeaways for CEOs and other leaders?

Rich Isenberg: The future is not humans versus AI; it’s humans *with* AI. As AI systems move from generating ideas to taking action, the real differentiator won’t be who adopts the technology the fastest. It will absolutely be who governs it the best. Agentic, sovereign, secure AI is not about slowing innovation. It’s about earning the right to scale it with trust, accountability, and control.

Rich Isenberg is a partner in McKinsey’s Atlanta office. **Lucia Rahilly** is the global editorial director and deputy publisher of McKinsey Global Publishing and is based in the New York office, and **Roberta Fusaro** is an editorial director in the Boston office.