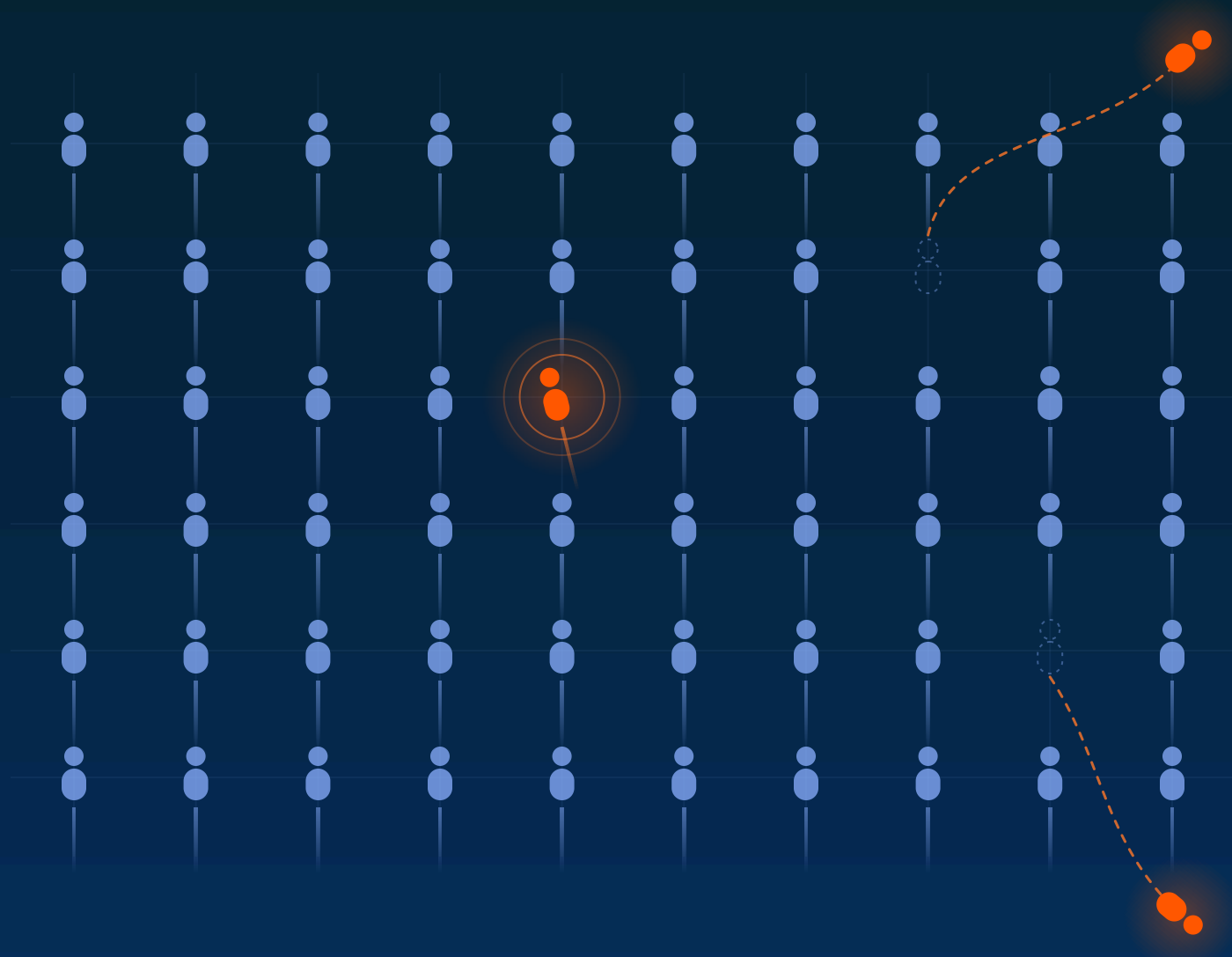




The Rogue Agent Report

Perspectives from AI practitioners on risk, unexpected behaviors, and what it takes to trust agents in production

The debate has moved on — to what agents are actually doing

Most organizations are no longer debating about whether to put AI agents into production, but are actually doing it. 85% of the teams across the study we conducted report having agents running in production, and nearly half of those agents are handling customer-facing or business-critical work. The challenge has now moved on into one focused on trust and reliability. Can we see what those agents are actually doing? Would anyone know if they did something unexpected? How do we mitigate our risk of improper access, breaches, and other security issues when humans are up against technology that can make its own decisions?

These fears around AI are legitimate, in part because technology is advancing and adoption is growing faster than governance can keep pace. Among enterprises we surveyed, two-thirds of teams deployed agents faster than they felt prepared to support because of internal pressures. The consequence of speed is less control; half of respondents say they have discovered an agent accessing data or systems they did not know it could reach.

That second finding is the subject of this report. As models become more sophisticated and organizations race to build their agentic capabilities, a lack of trust infrastructure to support these endeavors leaves enterprises vulnerable.

METHODOLOGY

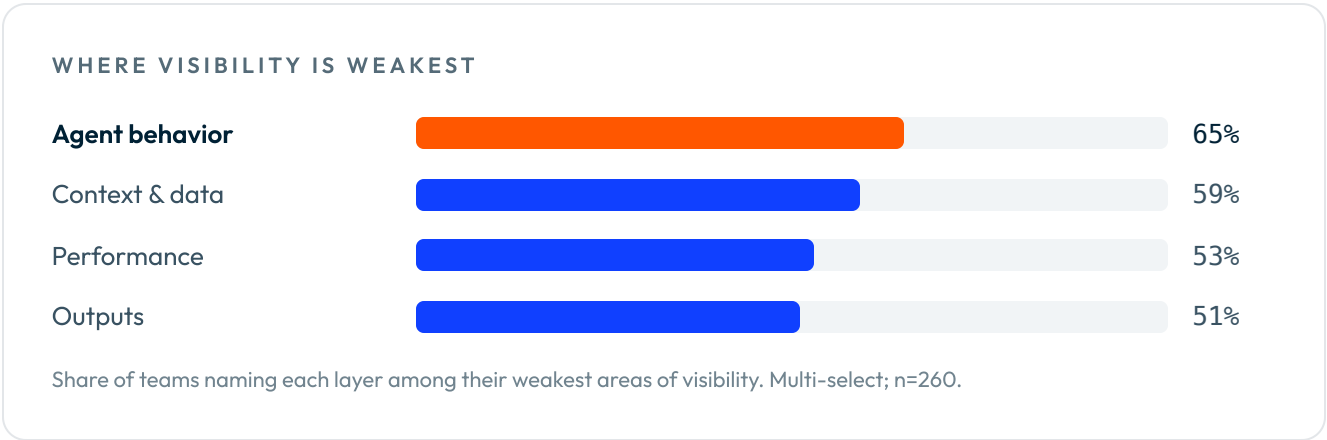
Online survey of 260 technology practitioners and engineering leaders at organizations with 1,000+ employees, conducted in early 2026. Respondents include engineers, developers and technical leads directly responsible for building and operating AI agent systems, alongside CTOs, VPs, Heads and Directors of Engineering. All responses complete.

Figures are reported on the full sample (n=260) unless otherwise stated.

01 Unauthorized data access by agents is widespread

49% of respondents say they have discovered an AI agent accessing data or systems they were not aware it could reach. Among organizations running agents in full production, this rises to 51%. The exposure grows with deployment depth, which is the opposite of what a maturity curve should look like.

The reason this shadow access goes unnoticed is visible elsewhere in the data. Only 41% of teams are fully confident they have visibility into all the data and tools their agents can access, meaning 59% are operating with acknowledged gaps in their own map of what their agents can reach. And when teams are asked where their visibility is weakest, the answer is specific:



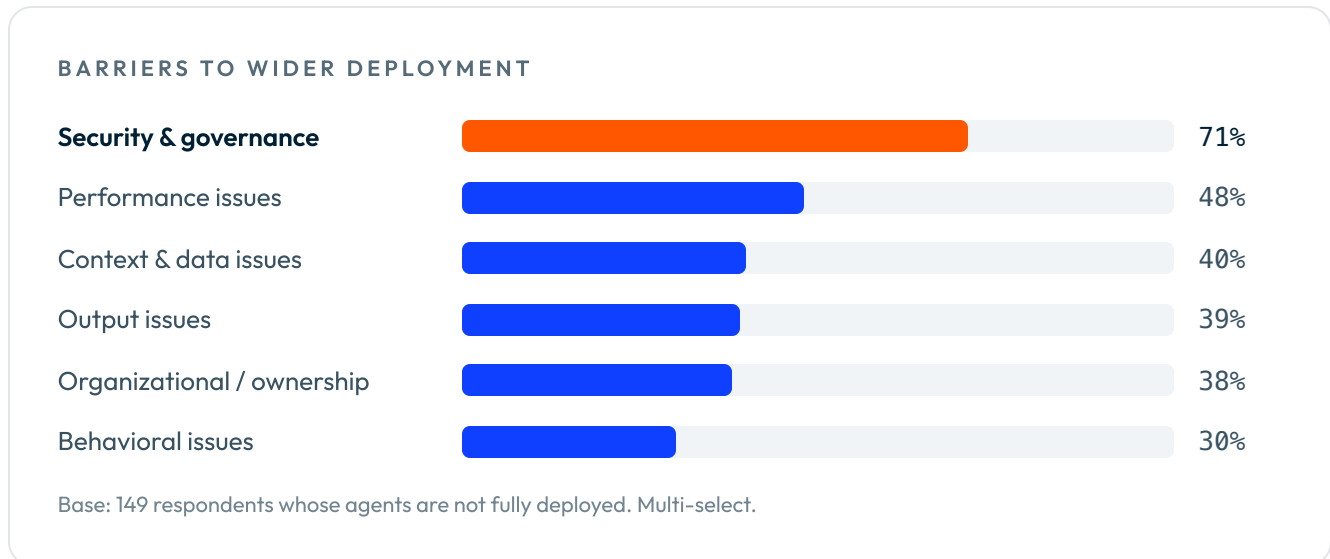
Behavior is the largest blind spot, yet it is precisely where unauthorized access happens. An agent reaching a system it should not touch is not an output problem or a latency problem. It is a question of what the agent decided to do, which tool it called, and what that call reached.

Half of organizations have already found an agent accessing data it should not have been able to access.

02 Security & governance are the main barriers to scaling AI

Enterprises are well aware of the shadow access risk, as many have experienced it, and it is the primary concern slowing deployment down. 47% of teams run agents in full production. Another 39% have agents confined to limited production; essentially internal or low-risk use cases only.

When teams are asked why agents are not more widely deployed, one answer dominates:



Security & governance is cited more than twice as often as agent behavior itself. Of the respondents who named that barrier, 29% already run agents in full production somewhere in their organization, and another 49% have agents working in limited production. This is not hesitancy from teams who have yet to start; it is organizations that have proven agents work, and are holding back on expansion due to security concerns.

The importance of security & governance also shows up when we asked teams what they believe is required for AI production-readiness. **Secure handling of data, credentials and tools is the single most-cited requirement at 69%**, ahead of monitoring and alerting (63%), performance guarantees (63%), and end-to-end traceability (56%).

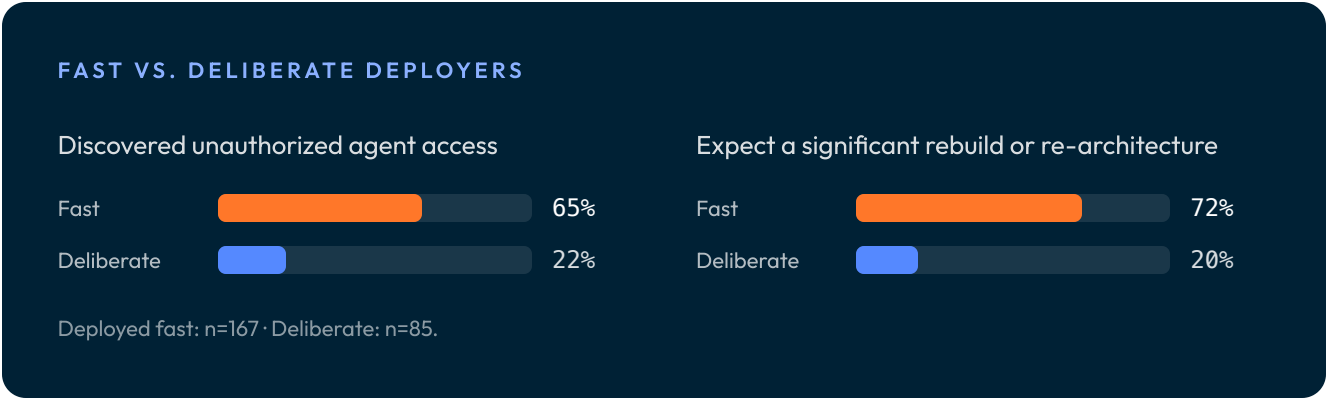
Despite this, only 50% of teams say their organization has documented, aligned criteria for what qualifies an agent as production-ready.

Security & governance is the top barrier to deploying more agents — cited by teams who already run agents in production.

03 Pressure to deploy quickly is a predictor for breaches and architectural rebuilds

So, is the caution warranted? The data says that it is, and it identifies organizational factors that further exacerbate the risk – namely, speed of deployment.

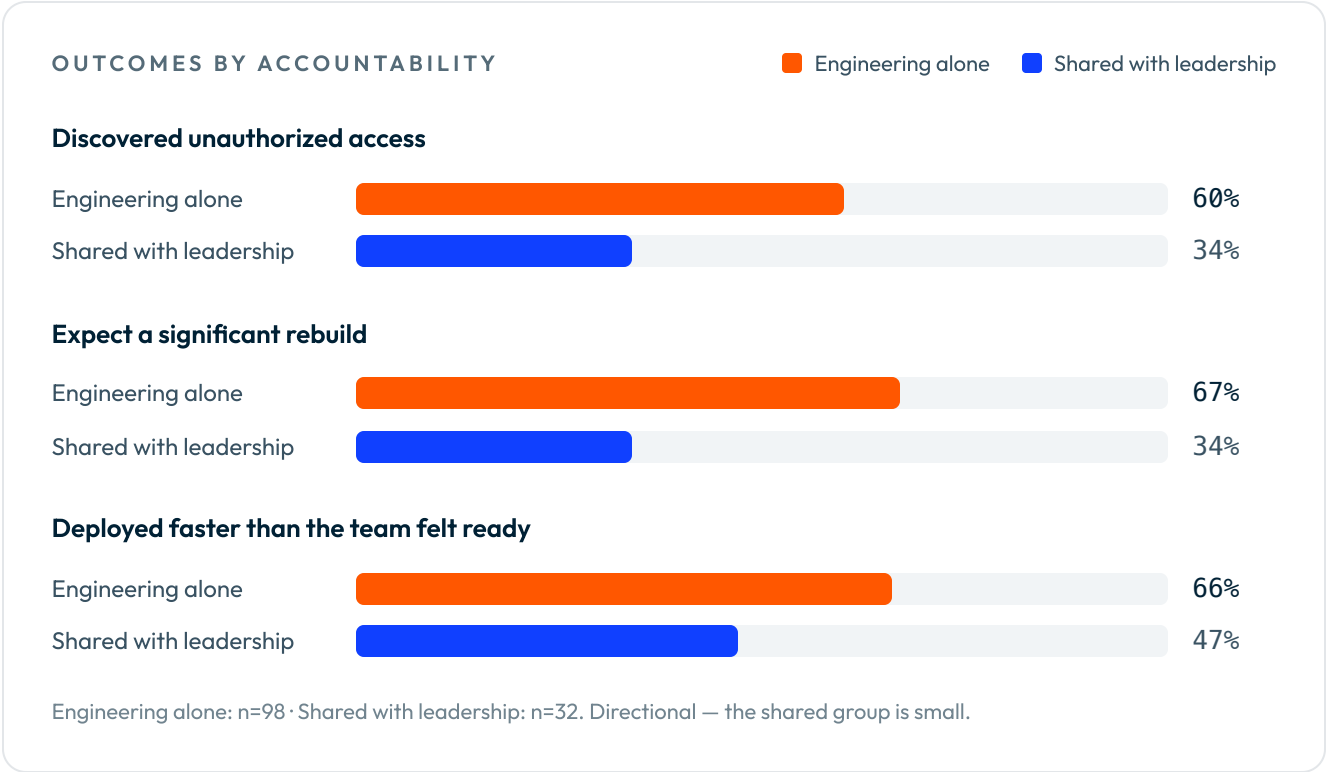
64% of teams reported deploying AI agents faster than they felt fully prepared to support due to business or competitive pressure. Comparing those teams against the ones who moved deliberately produces the sharpest divide in the research:



Teams who deployed quickly under pressure were 3x as likely to report incidences of unauthorized access by agents. They were also more than 3.5x as likely to say they expected to have to significantly rebuild or rearchitect the AI systems they stood up. The research shows us that speed does not merely create technical debt; it correlates with a greater risk of agents accessing things they should not, increasing security vulnerabilities and governance issues.

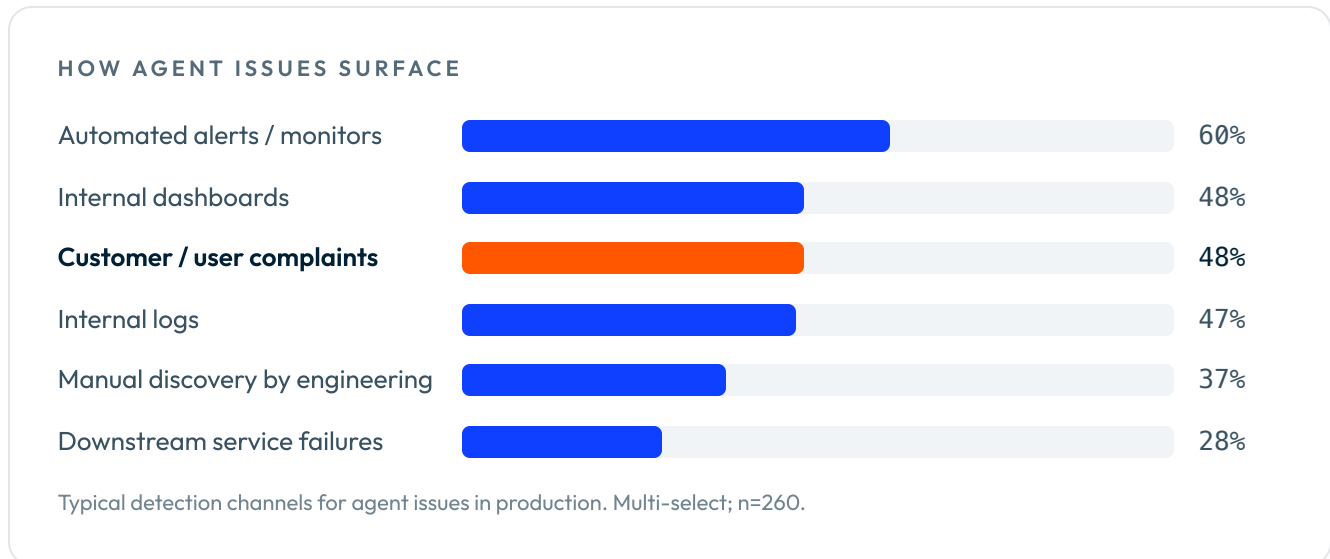
Shared accountability for AI reduces the risks

Speed is not the only structural factor; accountability is another highly influential piece of the puzzle. Where accountability for agent failures sits with the engineering team alone, outcomes are consistently worse than where it is shared between engineering and leadership. This is a promising finding: controlling for speed is unrealistic in today's ultra-competitive environment, but sharing the burden and creating a more equitable responsibility structure is an operational decision that can be made to support safer deployment.



04 Agent failures often emerge only when a customer raises the alert

Asked how agent issues in production are typically detected, nearly half of teams report that a customer or user complaint is what raises the red flag:



The picture is similar when teams describe how they know an agent is working correctly at all: 51% rely on manual spot checks and 45% on user feedback.

Then there is the harder question of what happens after an incident occurs. Only **48% of teams say an agent issue is easily traceable end-to-end**. The remainder: 32% can trace it, but only with significant manual effort; 19% have to stitch together multiple tools and logs; 2% describe it as difficult or nearly impossible.

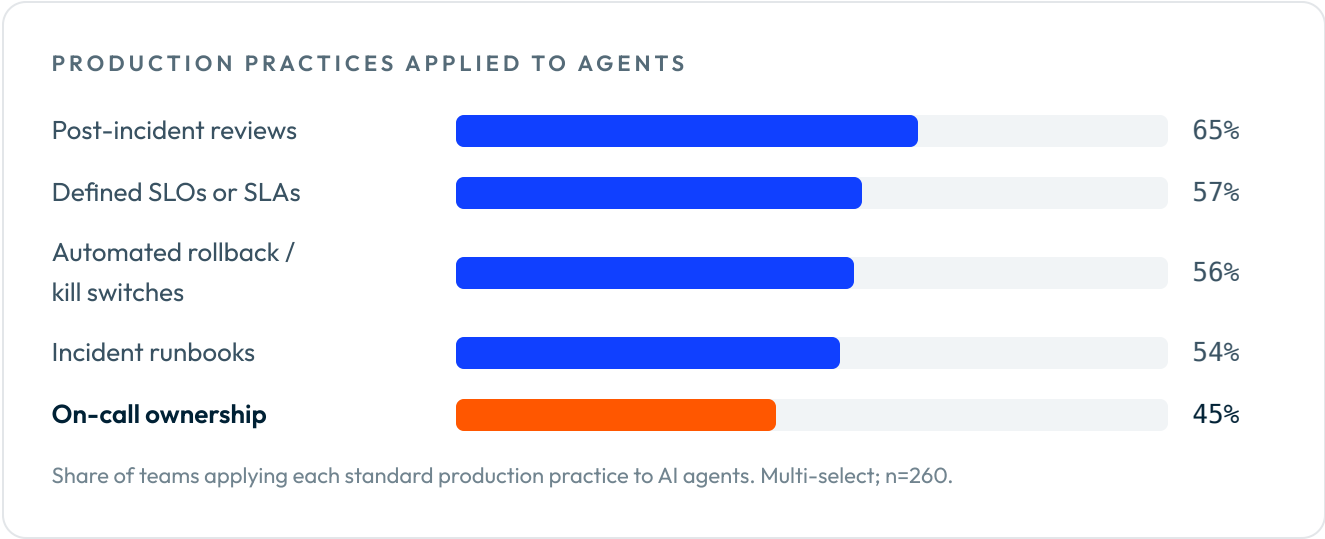
More than half of teams cannot reconstruct what an agent did without manual work. For an unauthorized-access incident, that is the difference between certainty and guesswork, and it introduces further risks around regulation and compliance requirements.

Nearly half of enterprises learn about agent issues via customer complaints.

05 Teams cannot roll back harmful agents quickly enough

The control mechanisms have not kept pace with the speed of deployment. **30% of teams we surveyed admitted they could not disable or roll back a harmful agent within minutes.**

The operational practices that make fast containment possible are unevenly adopted. Asked whether they treat AI agents like other production applications when it comes to standard production practices, teams report:



On-call ownership is the lowest-scoring practice in the entire survey. Fewer than half of teams have named a person responsible when a system that acts autonomously in production starts behaving unexpectedly.

41% rate their agentic incident response as less mature than their response process for other production systems. And among the teams who flagged agent behavior as a deployment blocker, 38% named the absence of guardrails or kill switches specifically.

Less than half of teams have on-call ownership for agents.
30% could not roll back a harmful agent within minutes.

A cautionary tale

Teams holding agents in limited production over security and governance concerns are responding to something real: half of their peers have already found an agent accessing data it should not have been able to touch. Those moving the fastest are the ones facing the highest number of incidents and are far more likely to have to rebuild their systems.

But the response to that risk has mostly been to slow down rather than to instrument. Behavior remains the least-visible layer at 65%. Customer complaints remain tied for second as a detection method. More than half of teams cannot trace an agent incident end-to-end without manual effort, and fewer than half have established a named owner to be accountable for an agent.

Each of those is a visibility problem before it is a security problem. An agent reaching unauthorized data is a security event, but finding out which data, when, and what the downstream consequences are across the rest of the stack is an observability challenge. For most teams, this is not something they can adequately address right now.

In order to truly achieve agent trust, teams will require a strong observability layer to complement the security practices they are putting in place to protect their organizations from the threat of rogue agents and nefarious uses of AI. Without a dual approach, enterprises leave themselves increasingly vulnerable as technology outpaces oversight, control mechanisms, and our ability to manage it.

Interested in learning more about agent trust, safety, and risk? Get in touch.

Trust your agents in production.

Request a demo →

montecarlo.ai/request-a-demo

ABOUT MONTE CARLO

Monte Carlo is the agent trust platform that intelligently monitors, troubleshoots, and improves agents and their underlying data — so enterprises can deploy reliable AI in production.