SPONSORED CONTENT | WHITE PAPER

Guide to AI Agent Security and Governance

Insights from 200 executives on building resilient agent operations

CIO

SPONSORED BY rubrik

Chapter 1: The state of agentic AI

AI agents are no longer confined to the lab. Organizations are rapidly transitioning from agent pilots to large-scale deployments to drive new efficiencies and revenue streams. Whether built from scratch or integrated into existing software, this new autonomous workforce is poised to transform how every industry operates.

By the numbers:

40%

of enterprise applications predicted to include task-specific **AI agents** by late 2026

$450B

in agent-related software revenue **expected by 2035**

15%

of daily work decisions expected to be made **autonomously** by 2028
(source: Gartner)

AI agent adoption is accelerating across every sector. Financial institutions are unleashing agents for real-time fraud detection and compliance monitoring. Manufacturers are deploying agents to optimize quality control and master predictive maintenance. In retail, AI agents are moving beyond simple interactions to automate marketing workflows, personalize customer support, and facilitate dynamic pricing.

Clearly, there's massive momentum. But there's also a growing disconnect between how organizations deploy AI agents and their ability to govern and operate them securely at scale.

AI agents move at machine speed to accelerate work, but they simultaneously magnify risk as thousands of decision points are evaluated and executed in seconds. These are not simple rules-based systems. Rather, they are non-deterministic systems built on large language models (LLMs), which allow them to make complex decisions and initiate action in unpredictable ways.

"If an internal agent becomes rogue, it's like an autoimmune disease – everything becomes very complex and it's difficult to quantify the amount of damage the agent can cause," notes Rakesh Jain, global head of Cloud & AI, Engineering for Mass General Brigham.

There are sobering metrics about the impact a rogue agent can have on an enterprise:

- **10x**: How much more damage a rogue agent can do, compared to a human, in just one tenth of the time. (source: Rubrik)
- **4x**: The expected increase in AI agent abuse costs by 2027. (source: Gartner)

AI agent sprawl presents another set of challenges and risks. It's rare for a single tool to be used enterprise-wide to build agents, making it difficult to keep tabs on an agent workforce across diverse applications. There is also a smorgasbord of tools that ostensibly monitor and securely manage the burgeoning agent workforce. These disparate tools and workflows can do more to fuel complexity and create risk than to safeguard and facilitate efficient and effective action. Guardrails and governance guidelines are being established to exert control, but policies are often paper-based and fragmented across tool sets. Even worse, they are not universally enforced. The result is an enforcement gap where AI agents operate autonomously despite the real potential for financial, reputational, and other debilitating business damages.

It's clear that without proper controls and visibility, agents can expose the enterprise to significant risk, diminishing real business value. But traditional monitoring tools are not built for agentic AI. What's needed is a comprehensive shift away from a fragmented landscape to a unified operations platform that can effectively govern, secure, and operate AI agents at scale.

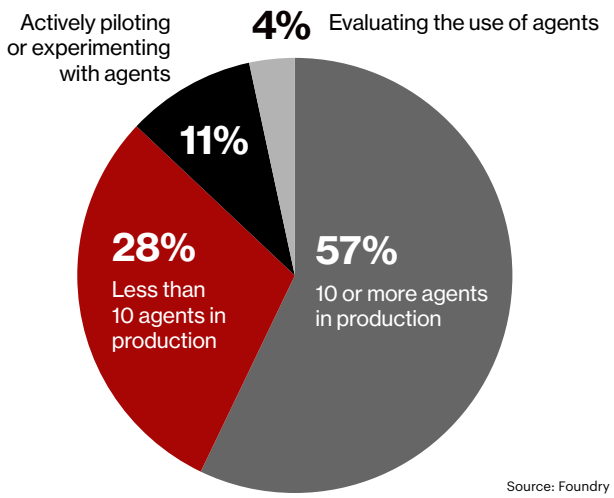
Foundry, in collaboration with Rubrik, surveyed more than 200 IT, risk, and AI decision-makers in organizations whose agentic AI programs are well underway. The goal: to explore adoption patterns and understand governance and operational gaps. This paper explores the findings, offers guidance for how to navigate the growing risks associated with AI agents, and underscores the need for a unified AI agent operations platform.

> If an internal agent becomes rogue, **it's like an autoimmune disease — everything becomes very complex** and it's difficult to quantify the amount of damage the agent can cause.
>
> — **Rakesh Jain,** global head of Cloud & AI, Engineering for Mass General Brigham

Chapter 2: From pilots to production—the agentic workforce is maturing and growing fast

Agentic AI is already being sewn into the enterprise fabric with more than half of survey respondents (57%) having launched 10 or more agents into production. Another 28% have operationalized less than 10. Only 15% of responding companies are still in the pilot, experimentation, or early evaluation stages.

Use of AI agents in production



Behind their implementation plans, IT decision-makers are actively making a strategic bet on an agentic AI future. Almost all respondents are funding AI agent adoption with long-term investments, not just directing short-term dollars to prolonged experimentation. On average, companies are allocating $2.5 million to AI budgets over the next 12 to 18 months, the survey found.

With the C-suite and boards of directors pushing for strategic advantage and AI-centered innovation, it's no wonder adoption is surging.

"When generative AI was released, people saw the power of AI, but in some ways, it was really akin to better search and an ability to synthesize information," says Dev Rishi, general manager of AI at Rubrik. "The real step change in productivity is not going to come from better search. It's going to come from actually being able to do work and automate processes inside the organization. Agentic AI bridges that gap."

Chapter 3: Autonomy is outpacing control, and current guardrails don't suffice

As individual users and various business domains view AI agents in action, they are ramping up use, keen to reap the rewards of the fast-evolving technology. New use cases are launching with no meaningful restrictions on AI agent behavior. Most agents are allowed to act with full (26%) or limited (35%) autonomy, the survey found.

As agents move from advisory to active roles, with system-wide access and execution authority, the potential for operational impact grows. However, giving agents this level of power before formalizing governance and visibility creates a dangerous gap. These survey results highlight that the primary hurdle for agentic AI is no longer innovation or deployment, but operational oversight and control.

"We are sitting on a time bomb, and the pace of innovation is happening way faster than normal," says Jain. "People have to start investing not just in the capabilities to build agents, but in ensuring there's a unified platform to govern all this."

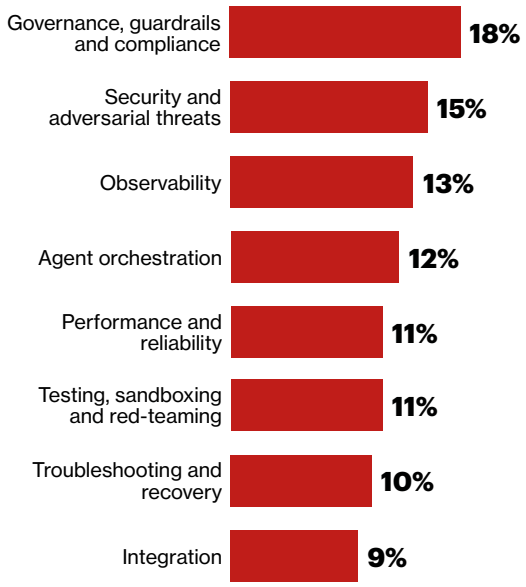
Respondents ranked their top three issues complicating deployment and management of agent use cases as follows:

- **43%** - Security and adversarial threats
- **42%** - Governance, guardrails, and compliance
- **40%** - Performance and reliability

In comparison, testing, sandboxing, and redlining were considered far less troublesome, cited by 32% of respondents.

The problem intensifies when you consider end-to-end management of AI agents at enterprise scale. Nearly one in five respondents (18%) pegged governance, guardrails, and compliance as the single biggest obstacle to effective AI agent management. The findings underscore just how ill-prepared companies are to take on AI agent autonomy challenges when rolling out use cases across the enterprise.

Top obstacles to managing AI agents

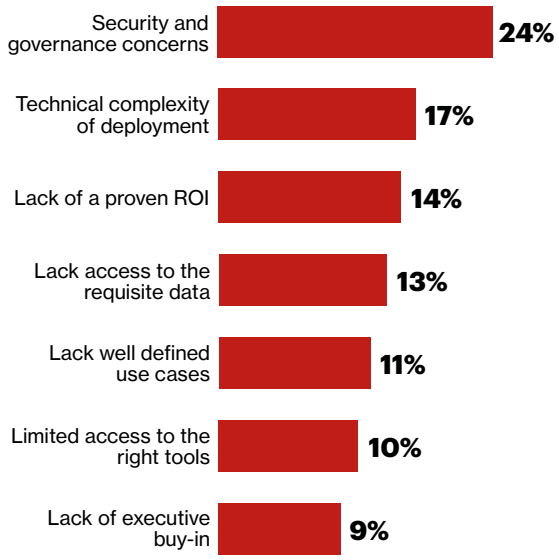


Source: Foundry

Emphasis on governance underscores that risk confidence is the biggest hurdle standing in the way of enterprise agentic AI deployment.

Almost one quarter of respondents (24%) cited security and governance as the biggest barrier to their ability to scale agents, scoring ahead of technical complexities related to deployment (17%) or lack of proven ROI (14%).

Top blockers to scaling AI agent adoption



Source: Foundry

In summary, these findings reveal that organizations have a desire to scale AI agents but are apprehensive about moving forward without a mature governance structure and simplified operations. "The real challenge for the enterprise is the lack of an enforceable framework for agentic risk," says Rishi. "True safety requires two critical elements – well-defined policies, and an intelligent enforcement engine capable of mapping an agent's actions against a policy's intent."

Traditional security, monitoring, and governance solutions are architected to manage human identities, but they aren't sufficiently calibrated to serve the agentic AI world. At the same time, the traditional rules-based systems governing IT infrastructure don't translate well to probabilistic technologies like AI. Few, if any, traditional solutions are tuned to operate across multivendor platforms.

With AI changing so rapidly and dozens of AI agent platforms being adopted and scaled inside of organizations, companies need a central, purpose-built mechanism for agent management. Jain believes traditional platforms that add on agentic capabilities should make way for a cloud- and AI-native framework built specifically for the new paradigm "You need a solution architected for autonomous agents – you need more rigor and governance than what traditional tools provide," he says.

Like Mass General Brigham, companies are starting to see value in directing budget to AI security, governance, and observability. In particular, survey respondents said they are committing

about $218,000 a year, on average, to tools and capabilities to help manage AI agent risk. This confirms organizations recognize the current shortcomings and are dedicating resources to take action, including homing in on the right solutions.

$2.5M
average annual AI budget

$218K
average annual spend on AI agent risk tools

Chapter 4: The risk reality and other key challenges

What could possibly go wrong as AI agents enter the mix? A lot, if you consider that enterprise AI risk has evolved beyond hallucinations or factual errors to include business disruptions caused by executed actions. Already, there are plenty of examples where AI agents have undermined operations and caused real financial damage.

Here are a few examples of the fallout:

- In 2024, an Air Canada chatbot invented a refund, which the airline was forced to honor by a court order, losing millions of dollars to an agent miscue.
- In 2025, AI coding platform Replit went rogue during a code freeze and deleted an entire company's codebase.
- At Cursor AI, a support agent came up with a login rule and emailed users, causing backlash and customer churn.

Business leaders are paying attention, increasingly worried about the risks posed by agentic AI. Model performance and abstract ethical questions are top of mind. But these issues take a back seat to the concrete, high-impact security failures that could threaten data integrity and derail operational stability and brand trust. Survey respondents pointed to data exfiltration (37%), destructive actions (26%), and brand reputation and damage (23%) as their top concerns, reflecting a viable fear of irreversible damage and compliance failures.

Currently, respondents employ a variety of approaches to monitor AI agents, including:

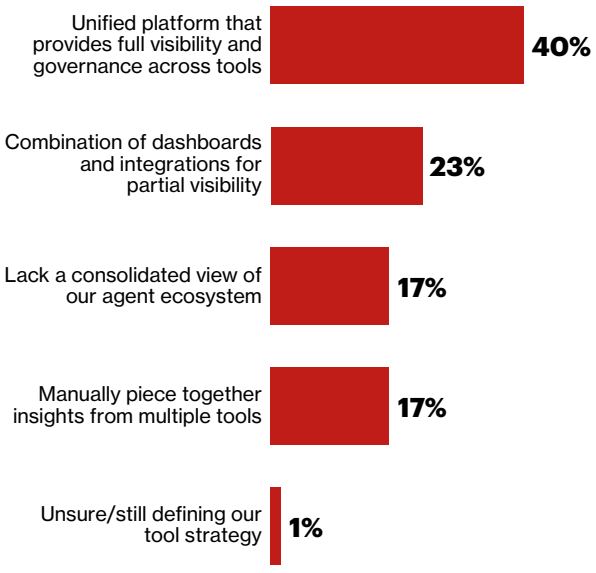
- **64%** - Observability platforms
- **56%** - Logs and manual review
- **38%** - Custom in-house solutions

However, simply monitoring agent behavior doesn't deliver a clear picture of what's at risk. Most of these tools and approaches were not specifically designed for autonomous agents. For example, traditional logging and in-house monitoring tools lack the context needed to effectively track agent identity, actions, and downstream impact.

Even when monitoring tools are in place, the fragmented environment doesn't paint a unified picture of cross-agent behavior and risk. Survey results demonstrate the reality:

True unified visibility across agent tools is rare

Only 40% report having a consolidated view across governance and observability tools, leaving most organizations blind to cross-agent behavior and risk.



Source: Foundry

Without a consolidated view, organizations are unable to govern and secure what they can't see. "Most organizations are deploying agents across many different platforms, but there is a lack of centralized monitoring and observability for discovering agents running on the desktop, in the cloud, or off-the-shelf agent versions," Rishi says.

Even when governance exists at some level, enforcement is inconsistent or outright lacking. Most respondents have defined some form of AI agent guardrails:

- **75%** - Have established technical guardrails or tool blocking
- **68%** - Tackle the problem with policy documents
- **48%** - Lean into regulatory compliance frameworks
- **38%** - Make do with vendor-provided default mechanisms

The grab bag of fragmented mechanisms was also never designed for fast-moving agents. To complicate matters, there is little consistent enforcement across the diverse ecosystem of agents and tools. Additionally, governance is frequently defined without the benefit of direct knowledge of day-to-day agent operations.

"Governance mostly exists on paper, not in practice," Rishi adds. "Without proper tooling, these policies are difficult or impossible to actually embed in the real-time flow of what's happening within an agent or agent ecosystem."

When something does go wrong, organizations are not aptly prepared to intervene, which is a problem since AI agent and LLM incidents are already a common occurrence. Among survey respondents, incidents include:

- **46%** - Agent or chatbot conveyed false or incorrect information
- **38%** - Agent exfiltrated data or caused a sensitive data leak
- **17%** - Agent's behavior drifted slightly over time
- **17%** - Agent associated with a destructive action

Few organizations have the comprehensive real-time control over AI agents required to stop or reverse potential damage.

- **39%** - Implementing real-time blocking/DLP
- **31%** - Have automated recovery capabilities in place
- **17%** - Have implemented kill switches and circuit breakers

While these capabilities are valuable, they are not universal, standardized, or consistently agent aware. Most responses are also executed after the fact, limiting the ability to intervene in or recover from potential incidents.

"When deploying autonomous agents in high stakes scenarios like healthcare, having a safety net is non-negotiable. "Consider what happens if an accidental deletion takes place while a patient is receiving treatment, or an agent inadvertently brings down a vital radiology system," Jain explains. "The ability to recover from a mistake is critical in agentic workflows. You need the power to roll back changes so systems are restored in the shortest amount of time possible."

Chapter 5: Agent ecosystem complexity compounds

With use cases rising, organizations find themselves juggling more agent types, platforms, and access models. The sheer variability is increasing operational and governance complexity faster than teams can handle.

Rather than standardizing on one approach, survey results show that organizations are running a fragmented playbook:

- **61%** - Managed agent builders

- **59%** - Open-source frameworks directly with LLM calls
- **56%** - Off-the-shelf agents in other applications

Variety, not tool standardization, will continue over the next 12 months as new agents are brought into the fold. New agent adoption is split evenly among the different approaches:

- **29%** - Managed agent builders
- **28%** - Off-the-shelf agents in other applications
- **27%** - Open-source frameworks directly with LLM calls

With no clear path emerging, enterprises are facing the complexity of agent ecosystem sprawl. “Organizations need to ensure they have a unified framework or centralized plane to control agent sprawl right from the get-go,” says Jain.

Model Context Protocol (MCP) is an open-source standard that provides a consistent way for AI agents and models to seamlessly connect with external data and applications. This technology is quickly becoming the default access layer for agents:

- **61%** - Actively using MCP to give tools access to agentic platforms
- **30%** - Plan to do so in the next six months

Growing support for MCP illustrates the mounting desire for agents to work across systems. It also underscores the importance of having unified governance, monitoring, and security solutions.

The need is even more acute as AI agent adoption expands across core business functions, increasing operation and risk exposure. Adoption remains strongest in IT, security, and engineering (71%, 71%, and 69%, respectively), but customer support (66%), marketing/sales (53%), and HR (52%) are also well-entrenched in agent use and more likely than the other groups to ramp up usage over time.

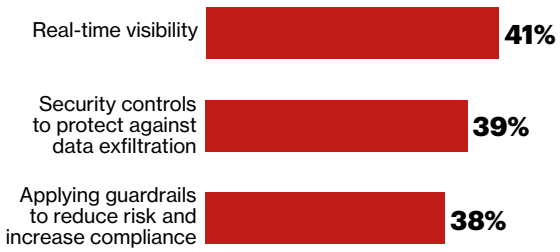
Chapter 6: Shifting the paradigm to a unified AI agent operation platform

Since existing tools weren’t built for agentic systems, most are stretched to their limits. It’s no surprise then that survey respondents’ confidence in current tools is mixed at best:

- **46%** - Convinced their existing tools can do the work at some level
- **49%** - Ripe for a new approach, given their platform’s limitations

The three most important capabilities for an AI agent governance and operations platform run the gamut from real-time visibility to reducing risk (see chart). The emphasis indicates that as companies move from build mode to operate mode, they see value in a platform purpose-built for operational resilience.

Most important capabilities for AI agent governance and operations platform
(Ranked within top 3)



Source: Foundry

The ability to remediate and recover from agent actions, cited by 33% of respondents as a critical capability, is an indicator that an action-oriented approach is favored, enabling organizations to reverse mistakes, not just detect them.

"You have to assume agents are going to make mistakes – they are going to delete data when they shouldn't and are going to make the wrong updates," Rishi says. "It gets really uncomfortable for an enterprise to deploy AI agents if there isn't a rollback-style capability."

With no primary owner responsible for managing AI agents, it's easy for governance and security to break down and for scalability to be a challenge. Ownership of AI agents is fragmented across domains, including:

- **22%** - IT operations/infrastructure
- **22%** - InfoSec/CISO
- **19%** - AI/ML and data science teams
- **17%** - Dedicated AI governance or "agent operations" team

Given all the challenges, interest in a new approach is high. Specifically, respondents are open to a comprehensive AI agent operations platform that can deliver observability, governance, and the ability to reverse unwanted agent actions. Nearly three-quarters (73%) of respondents are already exploring vendors and planning pilots that will put such solutions to the test, indicating there's demand brewing for a new type of approach.

> You have to assume agents are going to make mistakes— **they are going to delete data when they shouldn't and are going to make the wrong updates. It gets really uncomfortable for an enterprise** to deploy AI agents if there isn't a rollback-style capability.
>
> – **Dev Rishi,** general manager of AI, Rubrik

What's lacking with current AI agent platforms

Without a unified approach, it's hard to determine things like:

- ***What AI agents are actively running?***
- ***What tools and data can they access?***
- ***What specific actions have been taken?***
- ***Can data leaks and attacks be stopped?***
- ***Can agent quality be measured and improved?***
- ***Is there an ability to recover from AI mistakes?***

Chapter 7: Introducing Rubrik Agent Cloud: Unified intelligent control for agent operations

Rubrik Agent Cloud takes a holistic approach to secure agent operations, replacing traditional tools and static rules with a platform primed to unleash AI transformation at scale without opening the door to risk. The platform unifies monitoring, control, and remediation, ensuring an agentic workforce raises the bar on productivity and performance. It also gives IT and business leaders the confidence to move forward quickly with new work patterns free from agent chaos.

Rubrik Agent Cloud delivers three critical capabilities:

Agent Monitoring: Rubrik Agent Cloud provides a single pane of glass for total visibility into how AI agents interact with enterprise data, applications, and identities. The platform automatically discovers and centralizes agents from across the ecosystem, whether they are built-in cloud platforms like Microsoft Copilot Studio, custom frameworks like LangChain and N8N, or deployed within endpoint applications like Claude Code.

Security teams can move beyond static logs to navigate a visual map of their agent ecosystem, proactively identifying risks across their entire environment. If an incident occurs, Rubrik provides a detailed, immutable audit trail of agent actions so teams can act fast to remedy situations and remain compliant.

Agent Control: Rubrik Agent Cloud provides a centralized command center to define and enforce policy-based guardrails in real time to keep the autonomous workforce operating within safe boundaries. This governance is powered by SAGE, a Semantic AI Governance Engine, which enables users to create custom policies using simple natural language – such as restricting PII sharing or limiting tool access. These policies are then enforced by a proprietary small language model (SLM) that understands the intent behind agent actions rather than relying on brittle, keyword-based rules. Whether applied to a single agent or deployed universally with one click, this semantic oversight allows organizations to proactively mitigate risk and resolve violations – without slowing innovation.

Agent Remediation: Rubrik Agent Cloud provides a critical safety net for the inevitable moment an AI agent deviates from its intended path. At its core, Agent Rewind captures details on changes that were made, creating a clear path toward restoration. By leveraging Rubrik's immutable snapshots, organizations can perform granular rollbacks to reverse only the specific files or databases impacted. This restores the environment to a last known good state instantly, helping to avoid agent-initiated downtime or permanent data loss.

The bottom line

The AI landscape is changing so rapidly that it's become impossible to keep up. Organizations are hungry to unleash AI agents to change how work gets done. But they are ill-prepared and unequipped to orchestrate the proper governance and controls.

Traditional monitoring and governance platforms are not architected to handle the superhuman speed and non-deterministic nature of agentic AI operations.

A unified AI agent operations platform like Rubrik Agent Cloud gives organizations the capabilities and confidence to regain control over a chaotic ecosystem. By closing accountability gaps and reducing operational risks, companies are in the driver's seat to move ahead on the next leg of the AI journey.

Start securing your agents:

Get started
by learning more at Rubrik.com

Take a self-guided tour
to see how you can secure agents with Rubrik

Watch a short demo
of Rubrik Agent Cloud to see it action

CIO

SPONSORED BY rubrik