

Sponsored content | White paper

Closing the AI trust deficit

Here is what every CIO should know about explainable AI to accelerate adoption while mitigating the risks associated with access.



CIO

Sponsored by



A global financial services company deploys AI agents to automate identity and access requests across the business. Employees get faster access to the applications they need, and managers spend less time approving routine requests.

However, an audit uncovers that an AI agent has access to a financial reporting system. The organization can't explain how that access was approved or who's accountable for it.

This scenario is becoming more common as enterprises expand agentic AI deployments from isolated use cases into everyday business operations. [Research](#) shows that 51% of enterprises are actively deploying AI agents.

These agents can make decisions without human intervention, so businesses must put guardrails in place that keep humans in the loop before they can trust AI to handle their most sensitive tasks and data.

Considering the increasing volume of API keys, applications, [OAuth tokens](#), and other non-human identities (NHIs) that AI agents use to authenticate, guardrails are becoming an urgent issue. Research shows that:

- Some enterprises have about [45 machine identities](#) for every employee, rapidly increasing the number of NHIs that AI agents can amplify and manage.
- The security risk these non-human identities pose [has 67%](#) of enterprise leaders either concerned or very concerned about potential damage.

This is the AI trust deficit. Enterprises want the speed and efficiency of autonomous AI, but they need safeguards to keep agents from taking action that isn't aligned with their access policies, ethics, or other governance rules. They also need a consistent, readily available way to explain and govern AI actions.



"AI innovation has outpaced AI safety," says Mike Kiser, director of Strategy and Standards at [SailPoint](#). "The more this gap stretches, the more likely it is that organizations become ticking time bombs just waiting for an incident to happen."

An agentic-related incident could be a malicious action from an attacker like a prompt injection, or an agent taking an action on its own that violates data privacy.

Building AI trust and accountability starts by backing each decision an AI agent makes with clear, audit-ready context.

What audit-ready context looks like

Rather than rely on a black-box outcome, teams should be able to trace every decision back to the context that produced it. Boards, auditors, and regulators increasingly expect organizations to explain how AI agents make decisions and who's ultimately responsible for them.

Establishing audit-ready context can address these issues because it provides traceable knowledge from the AI agent's human owner to the specific policies triggered.

CIOs should be able to present context behind AI agent decisions in a way that non-technical audiences, such as boards of directors, will immediately understand. Here are questions the board might ask and the type of answer a CIO should be ready to provide:

Q: Who is accountable?

A: This decision was made by a finance access agent, which is owned by Sarah Chen, director of financial reporting.

Q: Why was this decision made?

A: The agent granted access because the request aligned with Policy FIN-204: Quarterly Financial Reporting Access and the user's role.

Q: What guardrails are in place?

A: This decision was determined to be low risk by our system and was therefore automated. The action was logged, and the agent's permissions are continuously monitored for any signs of behavioral drift.

Context should be easily accessible by anyone who needs to understand the reasons for that decision, whether it's an answer to a specific

question or to establish trust and safety. By mapping how agents, humans, and data are connected, teams can identify where sensitive data is at most risk and work towards closing any security gaps before they become an issue.

Continuous governance is critical

AI agents, business processes, and policies continue to evolve, which means that governance practices must evolve with them. As AI identities gain new responsibilities or interact with additional systems or apps, they can accumulate permissions that are no longer needed, increasing unnecessary risk. Here are some examples:

- **Cross-agent coordination:** Lacking strict inter-agent guardrails, a low-privilege inquiry chatbot can be manipulated via indirect prompt injection into calling a high-privilege administrative agent, tricking it into executing unauthorized system changes at machine speed.
- **Role changes:** An AI agent built for temporary data migration retains its original broad database privileges even after being demoted to a simple daily reporting task that only requires single-source read access.
- **Application lifecycle:** A financial automation agent updated to access a logistics system retains risky, lingering permissions to the legacy environment long after that logistics system is decommissioned and replaced.
- **Privileged access:** A service desk agent temporarily granted administrator-level permissions during a machine-learning training phase continues to run autonomously with those high-level privileges long after training is complete.

Ensuring evidence behind AI decisions

Every AI decision should be accompanied by:

- The AI agent's unique identity and its human owner
- Specific applications and data involved
- Exact policy rule(s) that governed the decision
- Precise timestamp
- Calculated risk score that justified automation or required human review
- Final decision outcome (approved, denied, escalated)

The privilege drift demonstrated in these examples makes it difficult to offer evidence that access still aligns with policy and the principle of least privilege, which states that users, apps, AI agents, and models should only have the minimum access and permissions needed to complete their tasks, and nothing more.

Traditional governance models, such as periodic reviews, can't keep up. The sheer velocity and volume of actions that an AI agent can take between periodic checks makes those reviews obsolete.

"Enterprises need AI governance capabilities that keep watch 24/7 and can stop an agent from taking additional actions if it starts doing something it shouldn't," Kiser says. "Without this oversight, an agent that started acting up at 3 a.m. could have done irreversible damage by the time the identity team woke up the next day."

The shift to continuous oversight helps enterprises identify and address risks.



Be prepared for governance questions

CIOs should be ready to answer these questions when asked by their CEOs and/or boards of directors regarding AI governance:

- ? How are we governing AI as it takes on more autonomous, important, and sensitive work across the business?**
- ? Who is accountable for every AI agent and the decisions it makes?**
- ? How do we know AI agents are making decisions that align with our policies and risk tolerance?**
- ? What evidence can we provide to demonstrate our AI governance is working?**
- ? Do we have the ability to shut down an AI agent and terminate its active sessions if it exhibits harmful or non-compliant behavior?**
- ? Do we have the visibility, governance, and protection needed to securely scale AI?**
- ? Are we prepared to take action on a rogue agent at any hour of the day to minimize adverse consequences?**

A framework for trusted AI governance

From an identity access management (IAM) perspective, AI agents act like digital employees. Every action must be authenticated and every decision involves authorization.

Enterprises successfully deploying agentic AI across their businesses share a common set of capabilities that span visibility, governance, and real-time response:

■ **Discovery and unified visibility:**

Unmanaged agents across SaaS platforms, browsers, and endpoints have the potential to cause serious damage. Businesses must continuously discover AI agents and other non-human identities, and understand their relationships to people, applications, entitlements, and data. All of this context should be consolidated into a single source of truth that is easily accessible for continuous management and auditability.

■ **Real-time governance and audit:**

The governance controls that apply to humans should also be extended to the non-human workforce, including access reviews and lifecycle management (creating and deleting machine identities used by AI agents). IT teams should apply policies consistently, provide audit-ready context for every identity decision, and continuously validate access as AI identities change.

■ **Proactive protection and response:**

Compromised AI agents act quickly, making manual responses inadequate and obsolete. Keeping up with the speed of AI agents requires an AI-powered governance and access management solution with capabilities such as continuous monitoring of the agents' work and the ability to quarantine or shut down an agent that's doing something problematic. In addition, proactive detection of behavioral drift and risk in real time, as well as automatic response around inappropriate access, allows teams to manage AI agents more effectively and protect against fast-moving security incidents.

Closing the AI trust deficit

Enterprises can build AI trust by making decisions visible, explainable, and accountable. With continuous identity governance and audit-ready context, organizations can deploy agentic AI quickly and more securely, advancing core business objectives while driving true innovation.

[Download the "CIO agentic AI governance checklist"](#) for a practical guide toward assessing your organization's AI governance readiness.

DISCLAIMER: THE INFORMATION CONTAINED IN THIS DOCUMENT IS FOR INFORMATIONAL PURPOSES ONLY, AND NOTHING CONVEYED IN THIS DOCUMENT IS INTENDED TO CONSTITUTE ANY FORM OF LEGAL ADVICE. SAILPOINT CANNOT GIVE SUCH ADVICE AND RECOMMENDS THAT YOU CONTACT LEGAL COUNSEL REGARDING APPLICABLE LEGAL ISSUES.