Illusion of control: Why securing AI agents challenge traditional cybersecurity models

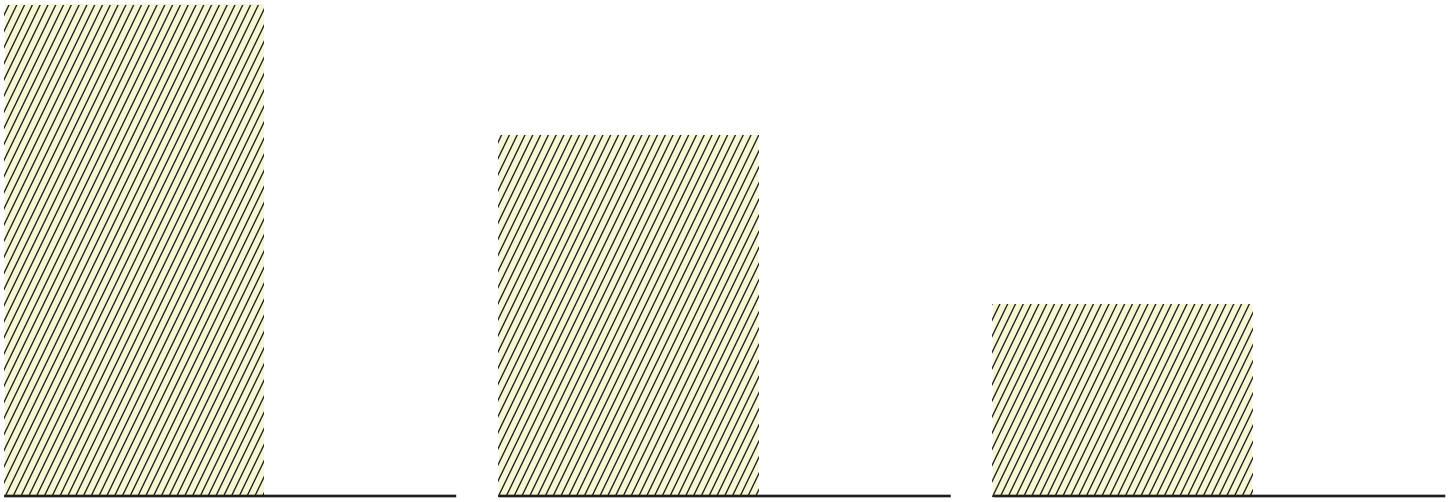
By Gal Zror, and Benny Porat

In this report

Enterprise security teams commonly focus on controlling AI agent conversations through prompt filters and testing edge cases to prevent unauthorized information access. While these measures matter, they miss the bigger picture: the real challenge is granting AI agents necessary permissions while minimizing risk exposure.

This isn't a new problem—it's the same fundamental challenge we've faced with human users for years. The solution involves applying proven strategies like just-in-time (JIT) enablement and ephemeral access, where permissions are granted only when required and strictly scoped to specific tasks. This zero standing privileges (ZSP) approach represents the evolution of AI agent security.

The urgency becomes clear when considering current adoption rates, **according to PwC's recent AI Agent Survey:**



88% of companies plan to increase AI budgets over the next 12 months

66% say they are seeing measurable productivity gains from AI agents

35% report broad organizational adoption of AI agents

With AI agents rapidly becoming integral to business operations, security teams must shift from reactive conversation monitoring to proactive permission management. The window for establishing robust AI agent governance frameworks is closing quickly as adoption accelerates across industries.

*Source: PwC AI Agent Survey https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html

Prompt security - Why prompt injection attacks miss the real risk

Security conferences repeatedly showcase the same demonstrations: researchers manipulating AI systems with clever prompts, bypassing safety measures, and extracting sensitive information.

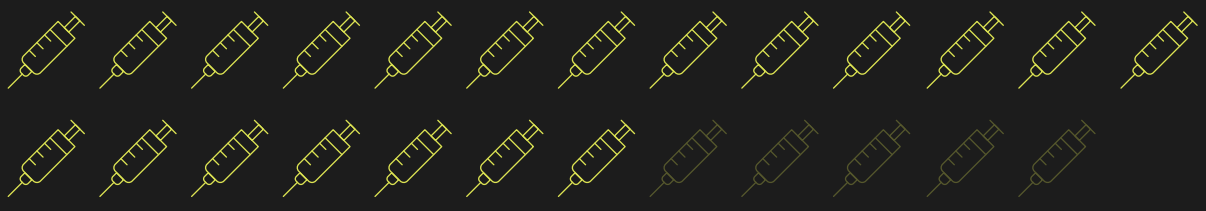
To back this up, recent NIST research concludes*:

that the average success rate for agent hijacking across a collection of five distinct injection tasks was

57%

Additionally after 25 repeated attempts on the same tasks, the average success rate for hijacking increased significantly to

80%



The audience takes notes. Vendors promise better filters. Everyone feels productive.

*Source: https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

However, prompt injection attacks are the cybersecurity equivalent of pickpocketing, visible, dramatic, and largely missing the point. While teams build increasingly sophisticated prompt defenses, their AI agents operate with standing access to systems that contain the actual valuable data.

Consider what happens after a prompt gets processed. The AI agent doesn't just think, it acts. It queries customer databases to answer support questions, pulls financial data to generate reports, and accesses employee records to handle human resources inquiries. Each action uses whatever permissions the agent was granted during setup, which typically far exceed what any task requires.

The manual oversight trap: Why human approvals slow AI security

Faced with these complexities, most organizations retreat to familiar territory: human oversight for everything important. AI agents need approval before accessing sensitive data. High risk tasks are banned entirely. The system gets reduced to handling only safe, routine queries that couldn't possibly cause harm.

This approach feels safe to executives who remember when automation meant scheduled scripts that ran overnight. It also eliminates most of the AI agents' business value. The efficiency gains, 24/7 availability, and ability to handle complex tasks at scale disappear into approval workflows that take longer than doing the work manually.

The manual approval model works until it doesn't. It collapses under high request volume, overwhelming human reviewers and slowing response times. And while teams struggle to keep up, competitors who've automated these processes begin to move faster, gaining a strategic edge. What feels like a safe, scalable solution is actually a temporary workaround disguised as a permanent policy.

To move beyond this illusion of safety, organizations need systems smart enough to make good decisions about AI agent access in real time, without requiring human intervention for routine tasks that fall within acceptable risk parameters.

> “AI agents need approval before accessing sensitive data. High risk tasks are banned entirely. ”

Zero standing privileges (ZSP) and agent interrogation: Enforcing just-in-time (JIT) AI access

AI agents have one advantage over other services or automation: they can be asked to explain their reasoning. Using a security agent gateway, **the agents must explain why access is needed before granting access to any system.** An agent's explanation can then be automatically verified against its defined role and current task. When a customer service agent requests access to financial data, the system can check whether that request makes sense given the specific customer inquiry being handled. If a data analysis agent suddenly needs human resources records, the system can flag the anomaly and either deny the request or escalate it for human review.

This interrogation happens in milliseconds, not minutes. The agent states its need, the system validates that need against the agent's purpose and current context, and access is either granted or denied. AI agents operate within much more predictable boundaries than human employees, who might have legitimate but unexpected reasons for unusual requests. The logical conclusion of this thinking is ZSP: no AI agent should have persistent access to systems it's not actively using. Every agent starts with no permissions and gets access only when needed, for specific tasks, and only for the minimum required time.

This marks a fundamental shift from the legacy model of granting broad, standing access to agents and trusting they'll use it appropriately. Instead, each action is gated by real-time authorization based on the current context and necessity. The agent must request access, justify its intent, receive a temporary credential, complete the task, and relinquish privileges, all in milliseconds. This dramatically reduces persistent attack surfaces. However, **agents must also be monitored for repeatedly trying to "fool" the system**, underscoring the need for agent behavioral analysis.

0

Default permissions

The ideal starting point for any AI agent in your system

100%

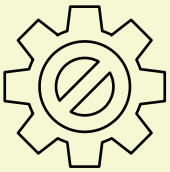
Justifications required

Agents must justify why they need every access request

2sec

Persistence time

After task completion, all privileges are immediately revoked



"This marks a fundamental shift from the legacy model of granting broad, standing access to agents and trusting they'll use it appropriately. Instead, each action is gated by real-time authorization based on the current context and necessity."

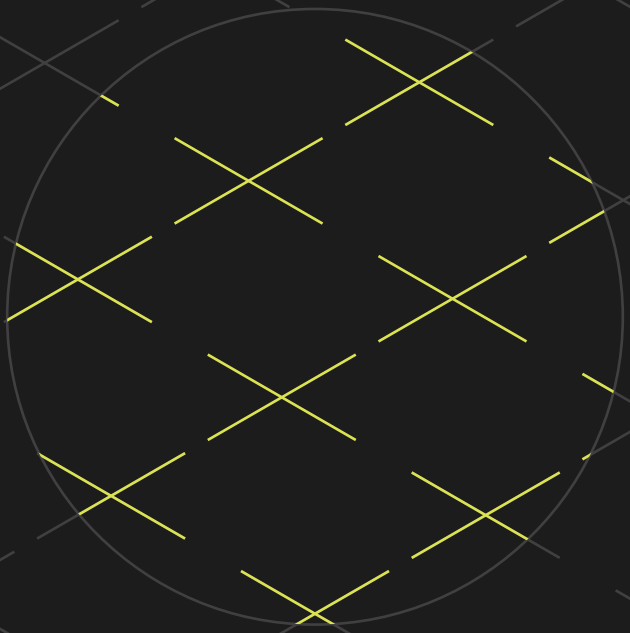
The shadow AI problem: Unmanaged agents and hidden security risks

Before any sophisticated access control can work, organizations need to solve a basic inventory problem: they need to know which AI agents they're running. This sounds simple until someone tries **actually to count them**.

AI agents often multiply in the dark corners of corporate infrastructure, **where visibility is low and oversight is often an afterthought**. Marketing deploys chatbots. Finance teams build data analysis tools. Developers integrate AI assistants into their workflows. Each team assumes their use case is unique and their security needs are minimal.

The result is "shadow AI" unmanaged agents operating with whatever credentials their creators happened to have available. These systems often lack proper identity management, credential rotation, or monitoring. They're the equivalent of unmarked server rooms in a building's security plan.

Each agent needs a verifiable identity and a secure way to get the credentials it needs. This isn't just about authentication; it's about establishing trust from creation through every action taken. The system **must distinguish between legitimate agents and malicious impostors**, between authorized and unauthorized deployments, and unauthorized experiments.



Dynamic AI security that actually works

The most important insight about AI agent security is that access needs change constantly. An agent processing customer service requests at 2 p.m on Tuesday needs different permissions than one handling emergency system recovery at 3 a.m on Sunday.

The security system must evaluate multiple factors in real time:

1

The specific task being performed

2

The sensitivity of the data being processed

3

The criticality of the systems involved

4

The current threat environment

These evaluations happen continuously, adjusting permissions up or down based on the immediate context.

This dynamic approach replaces the traditional model of static permissions—the same access level regardless of what's actually happening. Instead, agents get access based on immediate need, and lose it when the task is completed, and their risk profile is continuously monitored.

The calculations aren't complex, but they must be fast. AI agents work at machine speed, and security systems must keep pace without creating bottlenecks that eliminate the efficiency gains.

AI digital employees

Dynamic security is just one part of the picture. To support this kind of real-time, context-aware access, digital employees are emerging to handle the backend work:

- Documenting actions
- Analyzing patterns
- Tracking agent behavior
- Identifying efficiencies

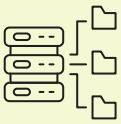
As agent spawning and collapsing accelerate to light speed, it's no longer humanly possible to track everything manually. You need AI managing AI. That's why digital employees, automated systems that augment your team, are becoming essential. They ensure that as agents perform tasks, the surrounding system remains secure, efficient, and constantly learning.

Building the right defense for AI agent security

What emerges isn't a single security product, but a comprehensive approach:

Discover every agent in the environment

Assign them verifiable identities



Use a unified gateway as a gatekeeper to enforce context-aware access controls

Eliminate standing privileges

Monitor behavior for signs of compromise

None of this requires breakthrough technology. The challenge is organizational, getting enterprises to treat AI agents as **fundamentally different entities** that need different security approaches than human users, or service accounts.

Organizations that solve this problem first will have a significant advantage. They'll deploy AI agents confidently, at scale, automating processes that competitors handle manually due to security concerns.

Watching for trouble: Detecting compromised AI agent behavior

Even with perfect access controls, AI agents can be compromised in subtle ways that don't immediately trigger alarms. For example, an agent might be manipulated to pursue objectives that serve an attacker's interests while appearing to function normally.

Behavioral analysis becomes crucial here. AI agents operate with consistent patterns. **They access the same data types, follow predictable workflows, and operate within narrow parameters.**

The system can flag an anomaly when:

An agent deviates from its established baseline

Requests unusual data

Works at unexpected times

Exhibits unfamiliar patterns

The current reality: Why AI agents need new security models now

AI agents aren't a future technology— **they're presently processing corporate data, making business decisions, and interacting with customers.** The question isn't whether to trust them, but how to trust them appropriately.

The current approach, focusing on prompt security while ignoring system-level vulnerabilities won't last. Manual approval processes and restricted use cases are temporary measures that will be abandoned as competitive pressure increases.

What's needed is the harder work of building security systems capable of managing the complexity of modern AI agents. This means moving beyond familiar approaches and embracing new **identity, access, and trust models.**

Most organizations already have AI agents operating in their infrastructure. The question is whether they can see them enough to manage them effectively.

This blog post was co-authored by CyberArk and Twine, reflecting our shared belief that securing AI agents begins with securing their identity.