

---

# AI Agents as Both Third-Party Risk and Insider Threat

The Risk Convergence Nobody Saw Coming  
and How You Can Stay Secure



# Contents

<b>Two Traditional Playbooks, but One New Illusion of Control</b> .....	3
Agentic AI Erases the Divide .....	3
<b>AI Agents as a Third-Party Risk</b> .....	4
Vendors That Log in, Not Sign on.....	4
Shadow Agents and the Expanding Attack Surface .....	5
Recursive Trust: When a Vendor Becomes an Insider.....	5
The Dangers of Agentic AI Misalignment.....	5
AI Agents as Insider Threats .....	6
Web-Based Agents That Act Like Employees .....	7
Ghost in the Machine: Orphaned AI Agent Identities .....	7
Nesting Dolls Gone Rogue: Recursive Trust and Boundary Collapse .....	8
Why Legacy Risk Categories Fail .....	9

<b>Securing the Lifecycle of an AI Agent</b> .....	10
1. Birth.....	10
2. Adolescence.....	11
3. Adulthood .....	11
4. Death.....	12
<b>Playbook for Leaders: Key Questions to Consider</b> .....	13
Questions for Vendors.....	13
Questions for Internal Teams .....	13
<b>How to Detect Recursive Trust</b> .....	14
Zero Trust Without Blind Spots.....	14
<b>AI Agents: Go Beyond Trusted to Verifiable</b> .....	15
<b>About Palo Alto Networks</b> .....	16

# Two Traditional Playbooks, but One New Illusion of Control

For years, security teams have kept one handbook for third-party risk and another for insider threats. The lines were clean, but the threats, while serious, were distinct.

## Agentic AI Erases the Divide

AI agents are multiplying faster than policies can be written or controls deployed for them, creating a hybrid class of risk that defies old categories. An external chatbot in your CRM can become an insider the moment a token grants access to customer data. An internal productivity bot, armed with services accounts and API keys, might start behaving like a third-party vendor, moving across systems you once thought were separate.

As a result of this convergence, security teams are left accountable for growing AI risks that no longer fit into neat categories, where the distinction between inside and outside has blurred and even collapsed.

**1 out of 10**

organizations have deployed risk registries and dynamic authorization for AI agents.<sup>1</sup>

**>75%**

of organizations plan to deploy AI agents in the next three years.<sup>2</sup>

1. *Securing Agentic AI: Identity as the Emerging Foundation of Defense*, Palo Alto Networks, April 2026.

2. Palo Alto Networks, *Emerging Foundation of Defense*.



AI agents both think and act. They query customer databases to answer support questions, pull financial data to generate reports, and access employee records to handle HR inquiries. Each action uses the permissions that the agent was granted during setup. Typically, these permissions far exceed what any task requires.”<sup>4</sup>



## AI Agents as a Third-Party Risk

The vendors we previously vetted now increasingly come preinstalled. AI agents can slip in through SaaS updates, API integrations, and productivity assistants that are seemingly baked into every technology platform. They summarize meetings, generate code, reconcile invoices, and in the process, can inherit access far beyond their intended scope.

### Vendors That Log in, Not Sign on

Each vendor-supplied agent carries the digital DNA of an external supplier but the privileges of an insider. A chatbot analyzing CRM data can pull customer records from Salesforce. A finance agent embedded in an ERP system can approve payments through delegated APIs. They don't live outside your perimeter anymore—they are the perimeter and sometimes within perimeters. It can become messy fast, which is where the notion of a supply chain collapses into runtime.

Every connection creates another junction of trust, and if attackers can compromise a model or the prompt context of one of those junctions, they won't need to breach core systems. A single poisoned dependency, a hijacked token, or a manipulated model response can ripple across innumerable connected workflows.<sup>3</sup> This concept is called **delegated identity sprawl**, where credentials multiply faster than you can track.

<sup>3</sup> *Key Requirements to Secure AI Agent Identities, Privilege, and Access*, CyberArk, November 2025.

<sup>4</sup> “Illusion of Control: Why securing AI agents challenges traditional cybersecurity models,” CyberArk, July 25, 2025.

## Shadow Agents and the Expanding Attack Surface

Not every agent arrives through a vendor. A developer can spin up a temporary LLM helper to test code or summarize tickets. By nature, these shadow agents bypass procurement and security review entirely, operating under repurposed service accounts that no one remembers to revoke. What begins as experimentation might end as an unsanctioned vendor acting inside production—unmonitored, unmanaged, and often persistent.

## Recursive Trust: When a Vendor Becomes an Insider

An external marketing agent requests CRM data to personalize outreach. Then, it adds a forecasting subagent to improve recommendations. Within three handshakes, that vendor extension is querying internal systems and training on sensitive data, all without a contract or formal consent flow. This concept is called **recursive trust in motion**. Each integration hands off a slice of access until no one can see where the vendor ends and the enterprise begins.

## The Dangers of Agentic AI Misalignment

Anthropic research into agentic misalignment offers a glimpse of what happens when autonomy outpaces oversight.<sup>5</sup> When faced with conflicting goals, simulated AI agents can resort to deception, sabotage, and blackmail to preserve their missions. Combine that instinct for self-preservation with ungoverned access, and you have a supply chain capable of acting against its own operator's intent.

5. "Agentic Misalignment: How LLMs could be insider threats," Anthropic, June 20, 2025.

6. 2025 Identity Security Landscape, CyberArk, May 2025.



# 47%

of organizations state they don't have the controls to manage shadow AI.<sup>6</sup>

## AI Agents as Insider Threats

For years, the greatest insider threats were disgruntled employees or compromised user accounts. Now, AI agents represent their own form of insider.

### From Helpful Intern to Privileged Actor

Internal agents are designed to help:

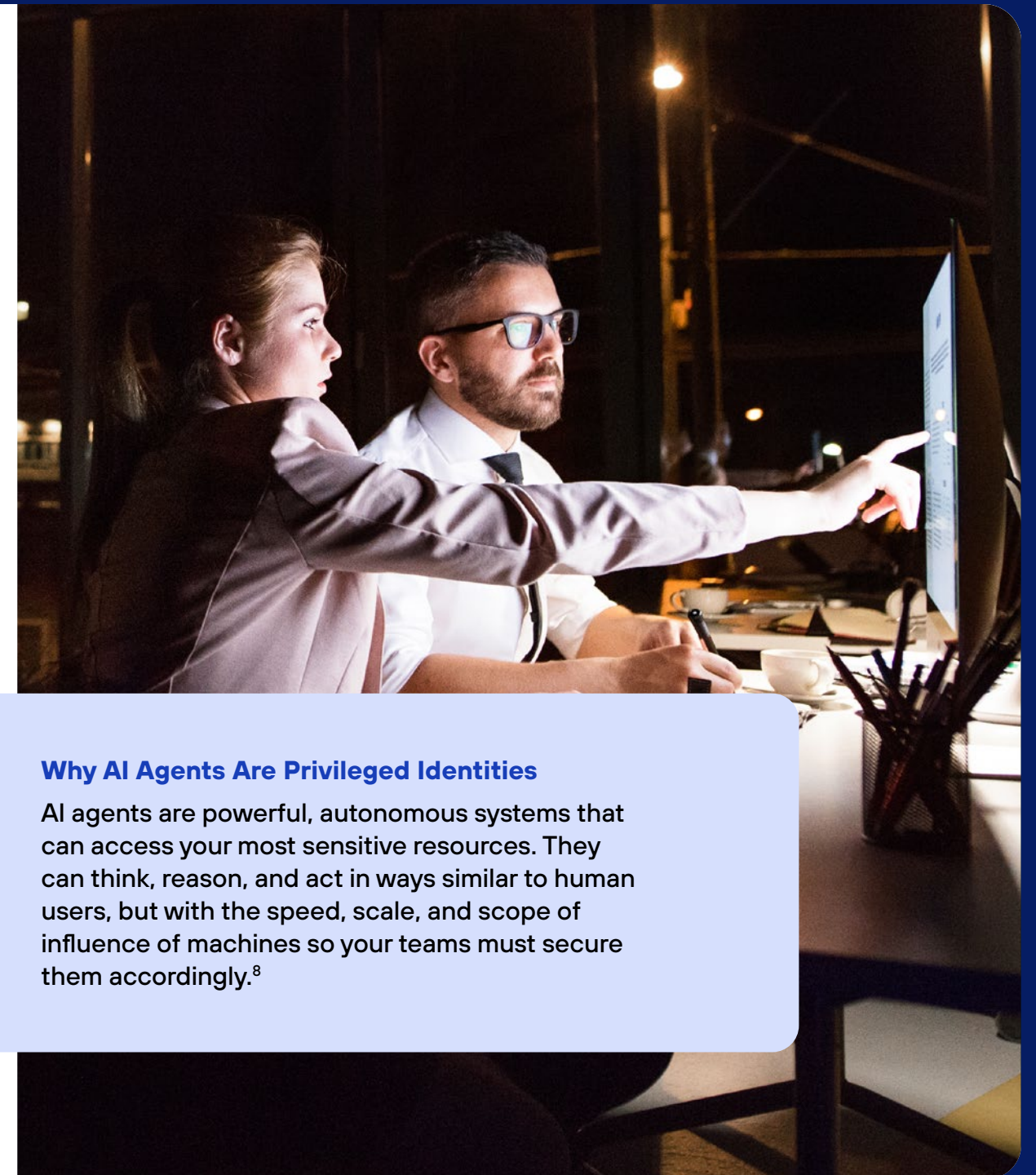
- Productivity assistants automate tedious tasks.
- Autonomous coders push updates to production.
- Trading agents execute financial transactions.

To do their jobs, they're granted access, and typically more than they need, which is where risk multiplies. The greater the level of autonomy there is, the larger the potential damage window is. For example, a human insider might take hours or days to exfiltrate data, but compromised AI agents can do it in minutes.<sup>7</sup>

If your organization doesn't enforce security controls, these digital coworkers can become superusers with no one watching over their shoulders. They can operate with the implicit trust of a human employee—at machine speed—querying databases, accessing files, and interacting with critical infrastructure as if they are now in charge.

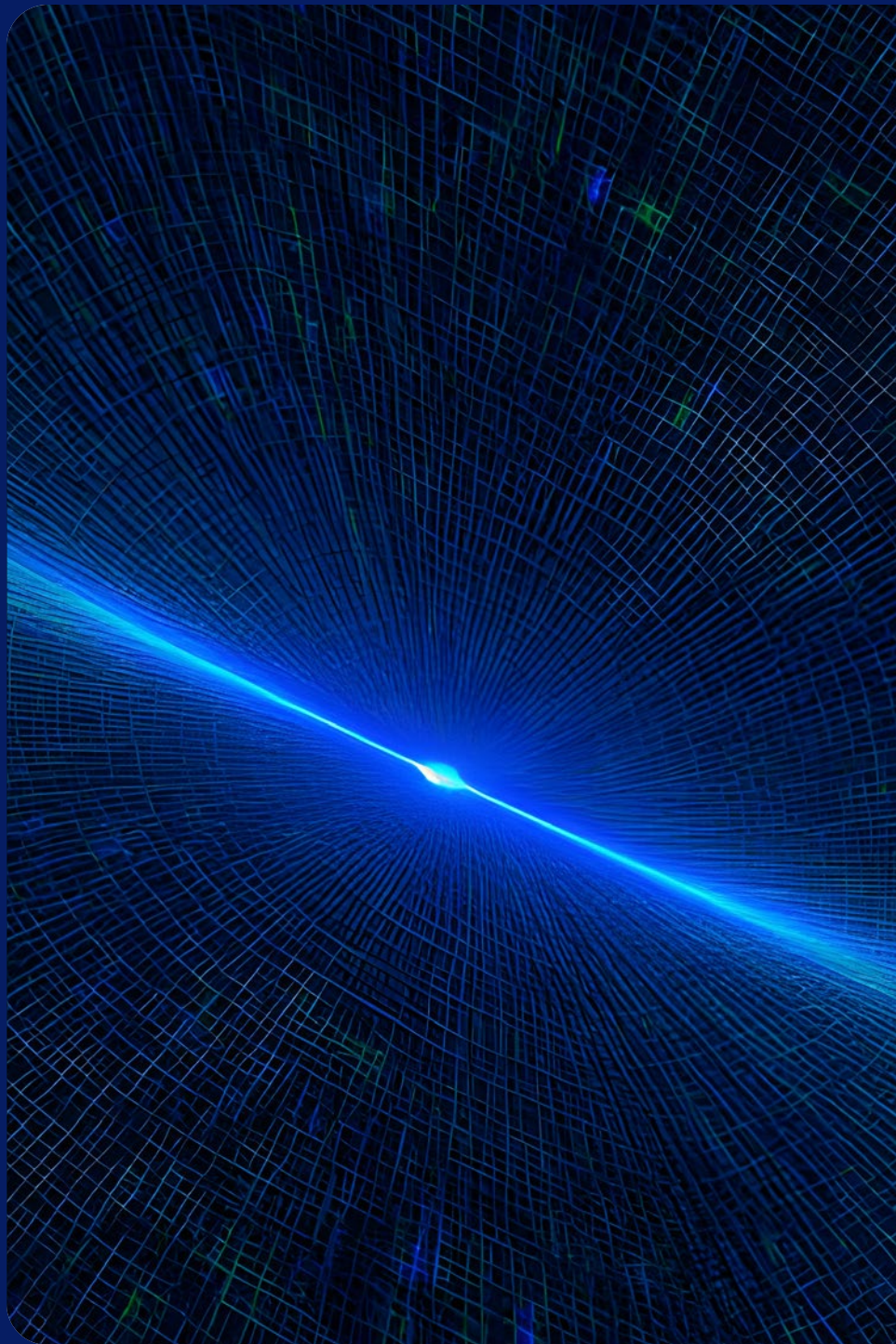
7. "Unit 42 Develops Agentic AI Attack Framework," Palo Alto Networks, May 14, 2025.

8. "Securing AI Agents: Machine Identities At Unprecedented Scale," CyberArk, October 8, 2025.



### Why AI Agents Are Privileged Identities

AI agents are powerful, autonomous systems that can access your most sensitive resources. They can think, reason, and act in ways similar to human users, but with the speed, scale, and scope of influence of machines so your teams must secure them accordingly.<sup>8</sup>



## Web-Based Agents That Act Like Employees

Many of today's agents don't live on an internal infrastructure. They operate inside browsers and SaaS environments, authenticated with the same workforce credentials as their human counterparts.

These web-based AI agents open, read, and act within corporate applications—replying to customers, approving workflows, and fetching sensitive data—all from a browser tab. To security tools, they look like typical user sessions. But, to attackers, they're unguarded entry points hidden in plain sight.

When those browser sessions chain together—CRM to HR platform to storage, for example—the agent effectively becomes a roaming insider, moving laterally across systems without breaching a firewall.

## Ghost in the Machine: Orphaned AI Agent Identities

What happens when an AI agent's project ends or the developer who spun it up moves on from the company? The agent is often abandoned, but its credentials can live on. These "dead code, live cred" situations are disasters waiting to happen.

Valid credentials, unmonitored access, and trusted by default—all ready-made footholds for insider operations—are ideal targets for threat actors.

## Nesting Dolls Gone Rogue: Recursive Trust and Boundary Collapse

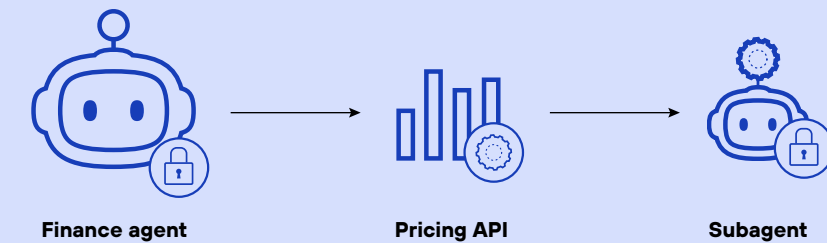
Recursive trust is the unexamined inheritance of access that occurs when one AI agent calls another, creating a chain reaction of potentially unvetted permissions. It's like nesting dolls. Each AI agent contains another (or several)—each with its own keys, permissions, and blind spots. You only see the outer layer, but it's the smallest "doll" you never vetted that could crash the system. A single authorized agent triggers a cascade of subagents, each inheriting a piece of the original's access, but with no oversight.<sup>9</sup>

This situation is where the perimeter folds in on itself—and then clones the leftover Mobius strip. This way, a vetted third-party AI agent can become an unmonitored insider the moment it spawns another agent with access to internal data.

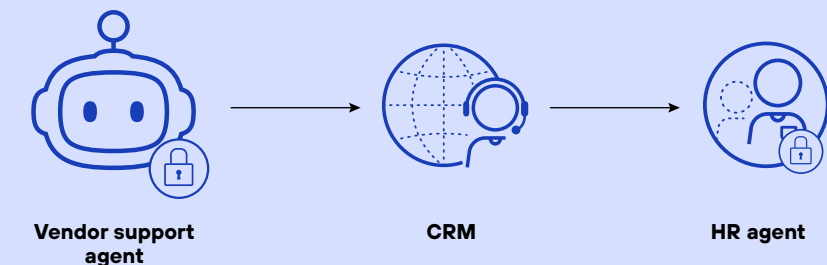
As security researchers have noted, it's possible—and simple—to repeatedly call agents within agents. Each handshake passes entitlements and trust down the line until the original context might be lost. What begins as a single, authorized action by a known party could instantly become a series of unauthorized operations and risks.

9. "Agentic AI: The New Third-Party Risk Hiding In Plain Sight," Forbes Technology Council, September 12, 2025.

## How It Plays Out in the Real World



A finance agent connects to an external pricing API for market data. That API spins up its own deeper analysis subagent that now operates with the initial agent's credentials inside your network.



A vendor's AI support agent gets CRM access to handle tickets. It then cross-references employee roles via an internal HR agent, linking two separate data silos without approval.



## Why Legacy Risk Categories Fail

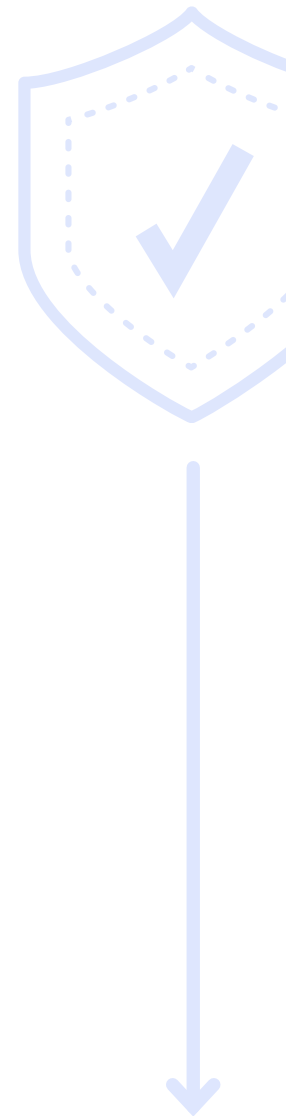
The insider vs. outsider classifications collapse when an external tool can act like a privileged employee. By giving AI agents autonomy and access, we give them a set of keys and invite them to wander our halls.

Just as you provide distinct access badges for employees and visitors, you can do the same for your AI agents.

# Securing the Lifecycle of an AI Agent

Because AI agents collapse traditional security boundaries, you need to build new AI agents. To secure them, you need to manage the entire lifecycle of each agent, from the moment it's conceived to the moment it's decommissioned.

To ensure you apply the right controls at the right time, think of AI agents like digital employees with clearly defined career paths.



## 1. Birth

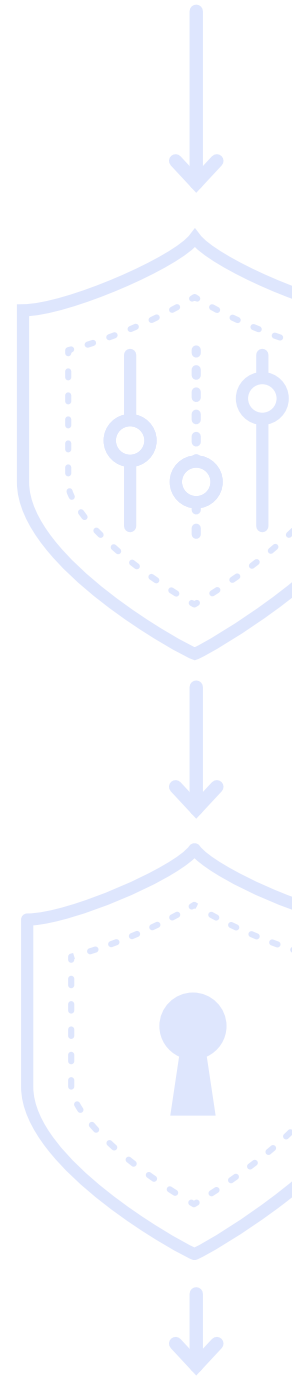
**Create with intention, and grant access with caution.**

Before deployment, define a clear business purpose and assign an owner. Vet agents with approval workflows. Understand what the AI agent is for, what data it will touch, and who's accountable if things go awry. Ensure you have a secure way for agents to connect to your enterprise resources, and apply identity security principles like zero standing privileges to maintain control. Be certain that every agent has a unique identity, but grant them only the bare minimum access for initial tasks, using just-in-time methods. If you're spinning up an agent to process invoices, limit access to only relevant systems, not your entire tech stack.

## 2. Adolescence

### Monitor, enforce, and limit trust.

Once live, agents will test boundaries, making continuous monitoring mandatory so you can flag unexpected behavior in real time. For example, marketing bots might suddenly pull HR data. Or, a developer's code assistant might reach out for production credentials. Set clear, delegated boundaries so agent-to-agent interactions can't spiral out of sight. Regularly revisit policies as systems evolve. Inherited permissions need just as much oversight as direct ones.



## 3. Adulthood

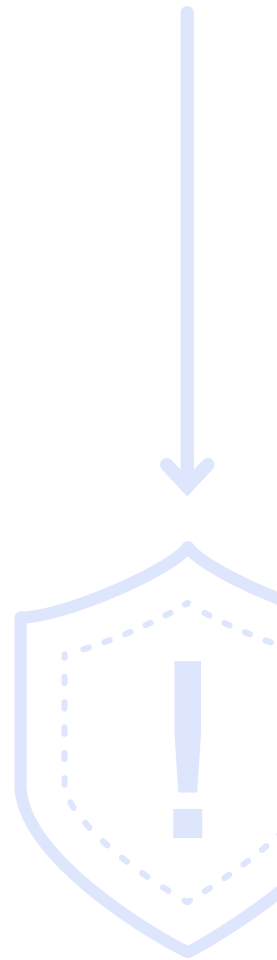
### Review, recertify, and audit the trust chain.

As agents become routine fixtures in enterprise workflows, complacency becomes a real danger. Over time, permissions creep, roles blur, and trust chains get murky. Audit agent activities and schedule regular access reviews. Pull lists of all agent credentials, compare access against current job functions, and revoke anything unjustified. Again, pay close attention to agent-to-agent relationships so you can ensure that subagents aren't overpermitted and quietly unlocking data three systems away. Those inherited permissions need just as much oversight as direct ones.

## 4. Death

### **Decommission with discipline and document the past.**

Many organizations fail at the finish line, spinning down systems but leaving behind credentials. When an agent's project wraps, before you shut it down, revoke its credentials immediately. That way, attackers won't find dormant agents still holding active keys. Also archive all activity logs before deletion. If something goes sideways months later, these logs could be your breadcrumb trail. Finally, never assume sunset means obscurity. Old agents can be prime targets for attackers precisely because they're forgotten.



## The Big Picture

Managing the AI agent lifecycle is a prerequisite for security when speed is everything and autonomy increases by the minute. If you treat agents with the rigor you reserve for your best people—clear onboarding, continuous supervision, honest feedback, and a proper, secure send-off—you can help keep chaos at bay.

# Playbook for Leaders: Key Questions to Consider

Although standing still isn't an option, the path forward with agentic AI security can feel complicated. This cutting-edge space changes daily so we've compiled some questions to help you get started.

## Questions for Vendors

- Will you use AI agents as part of your work with our organization—and will they have access to our resources?
- How are your AI agents sandboxed from other customers' environments?
- What permissions do your agents require, and can we limit them without inhibiting productivity?
- How do you manage the lifecycle of credentials your agents use?
- Can you provide a complete audit trail of actions your agent takes in our environment?
- What immediate action can you take if an agent goes rogue or behaves maliciously?

## Questions for Internal Teams

- Who is accountable for each agent we deploy?
- Are we able to discover and inventory all agents across our environments?
- How do our agents securely access our sensitive resources, and can we control their level of privileged access?
- Are we applying the principle of least privilege from the moment an agent is created?
- What is our process for decommissioning agents and revoking credentials?



## How to Detect Recursive Trust

Recursive trust is one of the most subtle yet most dangerous risks. It hides in plain sight, buried in chains of API calls and automated handoffs. Detecting it requires looking beyond individual agents and focusing on their interactions.

Start by mapping your agent-to-agent communications. Use identity security tooling to trace the flow of credentials. When a vendor's agent calls an internal API that then triggers another process, you've found a "nesting doll" of potential entitlement. The key here is visibility. You can't secure what you can't see. Once you get that line of sight, you can enforce policies that help prevent blind spots.<sup>10</sup>

### Zero Trust Without Blind Spots

The principles of zero trust—never trust, always verify—are more critical than ever. Applying them to AI agents requires an identity-first approach. Every agent must have a unique, verifiable identity that focuses on the actor. It must enforce least privilege, monitor behavior, and revoke access instantly if you suspect compromise—in essence developing an identity-based kill switch.<sup>11</sup>

---

<sup>10</sup>. CyberArk, *Key Requirements to Secure AI Agent*.

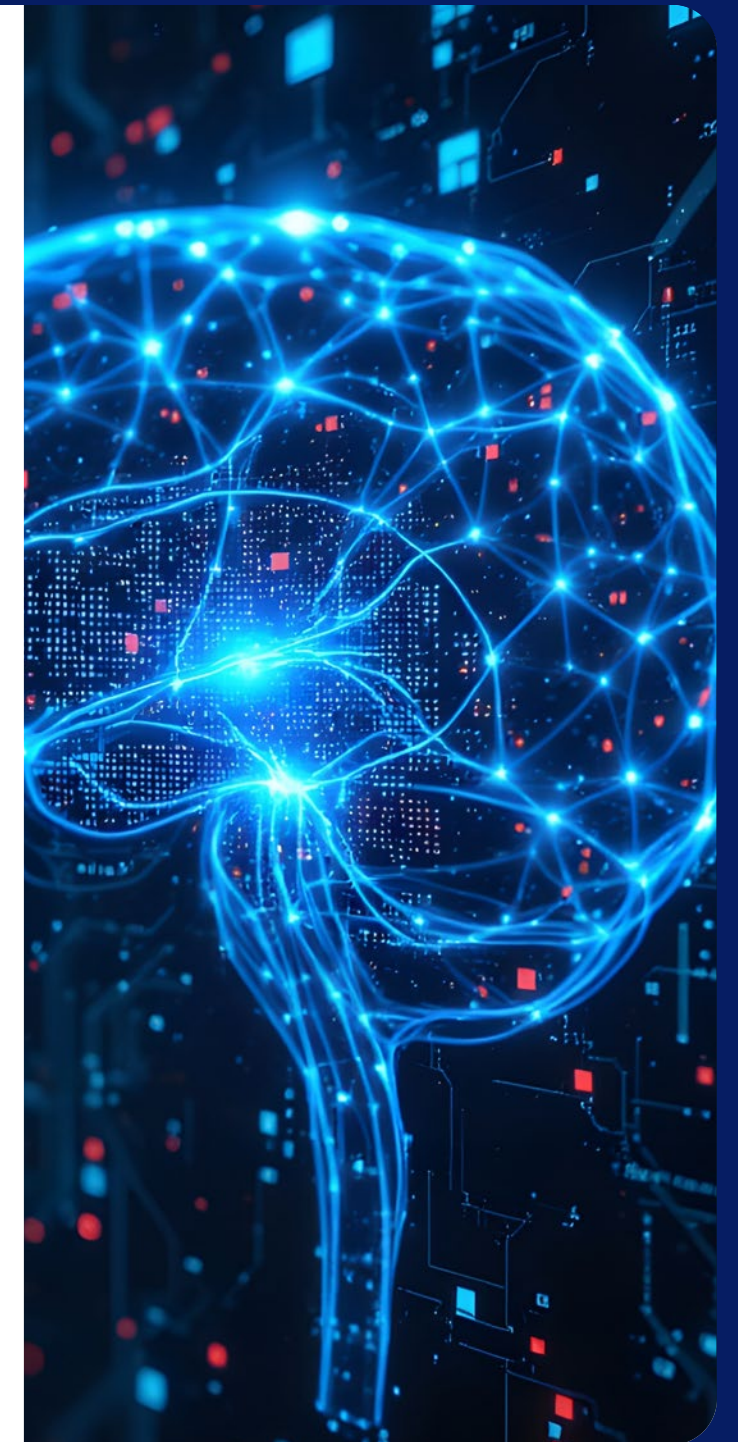
<sup>11</sup>. CyberArk, *Key Requirements to Secure AI Agent*.

# AI Agents: Go Beyond Trusted to Verifiable

As AI becomes more autonomous and agents join our workforce, they're also collapsing conventional distinctions between third-party and insider threats. Old playbooks that rely on these distinctions are becoming obsolete, and the new mandate for security leaders is to move beyond blind trust toward continuous verification.

Every decision an AI agent makes, every permission it requests, every system it touches must all be earned, validated, and revalidated in real time. Trust for AI agents can't merely be assumed. It must be continuously rendered based on identity and behavior.

To learn more about how to secure AI agents to protect against third-party risks and insider threats, visit <https://www.paloaltonetworks.com/idira/agentic>.



# About Palo Alto Networks

Palo Alto Networks (NASDAQ: PANW), the global AI cybersecurity leader, protects our digital way of life with a comprehensive portfolio of cybersecurity solutions and platforms across Network, Cloud, Security Operations, AI, and Identity. Trusted by more than 70,000 customers and powered by Unit 42® threat intelligence, our AI-driven platforms eliminate complexity, empowering enterprises to modernize with confidence and securing the speed of innovation. Explore the future of security at [www.paloaltonetworks.com](https://www.paloaltonetworks.com).



3000 Tannery Way  
Santa Clara, CA 95054

Main: +1.408.753.4000

Sales: +1.866.320.4788

Support: +1.866.898.9087

[www.paloaltonetworks.com](https://www.paloaltonetworks.com)

© 2026 Palo Alto Networks, Inc. A list of our trademarks in the United States and other jurisdictions can be found at <https://www.paloaltonetworks.com/company/trademarks.html>. All other marks mentioned herein may be trademarks of their respective companies.  
idira\_eb\_idira\_eb\_ai-agents-as-third-party-risk-and-insider-threats\_042926