servicenow

INNOVATION BRIEF

# Trust as Infrastructure

## Building secure systems in an agentic era


WRITTEN BY

**Dave Wright**
Chief Innovation Officer


The Innovation Office

# Trust as Infrastructure

## Building secure systems in an agentic era


## Summary

- Agentic AI is forcing enterprises to reckon with their existing trust and security infrastructure.
- Organizations can embed trust into their AI architecture by focusing on visibility and control, building robust monitoring and identity infrastructure, prioritizing active human oversight and cross-functional governance councils, and deploying advanced AI security capabilities.
- Unlike human users, AI agents operate autonomously across systems and can modify their own behavior, requiring fundamentally different security approaches.
- Creating a solid trust architecture today will enable organizations to deploy AI capabilities that competitors cannot risk, directly translating to market dominance. Those who wait will be left scrambling to retrofit security after breaches occur.


Trust is "the new currency"—or so goes the latest popular refrain.[1] At a macroeconomic level, trust improvements correlate directly with GDP growth: A 10% increase in the share of trusting people raises GDP growth 0.5%.[2] The same applies to corporations and brands, whose trustworthiness directly impacts their market position and customer selection.

In enterprise IT, trust has traditionally taken on a more specific and technical meaning: the confidence that systems will behave reliably, securely, and as intended. Thanks to agentic AI, this definition is evolving. As companies deploy AI agents that access sensitive intellectual property, make autonomous decisions affecting reputation, and operate across complex regulatory landscapes, trust is neither just a perception nor a technical specification, but something much more comprehensive and crucial. In the agentic era, trust requires both reliability (ensuring AI agents execute within defined boundaries) and intent (continuously verifying they act with transparency into their decisions and oversight of their impact).

Organizations with demonstrable trust, verified through governance frameworks, security controls, and transparent operations, can deploy AI agents that their competitors cannot risk. Those without it face compounding disadvantages: restricted or uncontrolled use cases, regulatory fines, and ultimately, lost market share.

Gone are the days when security was retrofitted in an enterprise system. In the agentic era, trust

must be a core design principle, embedded from inception.

CISOs and CIOs are not just managing static applications and human users. They are now governing entities that can modify themselves and make decisions at machine speed, requiring a fundamental redesign of security architecture. For CEOs, this means that agentic AI security isn't just an IT problem to delegate. It's a strategic risk and opportunity that will determine whether they capture market share or must explain breaches to the board, pay out hefty fines, and lose customers.

> **Gone are the days when security was retrofitted in an enterprise system. In the agentic era, trust must be a core design principle, embedded from inception.**

Globally, while people increasingly believe that the benefits of AI outweigh the drawbacks, confidence in AI companies' protection of personal data fell from 50% to 47% in a single year.[3] According to a 2025 report by IBM, one out of six security breaches already involves AI.[4] It's alarming, then, that 97% of businesses that reported an AI-related breach lacked proper AI access controls.[5] Gartner®[6] predicts that "By 2028, 25% of enterprise breaches will be traced back to AI agent abuse, from both external and malicious internal actors."[7] Add deepfakes and other AI-driven attacks to the mix, and businesses are now facing a significant and rising threat that most are not ready for.

The good news? Organizations can fight fire with fire. On average, companies that successfully integrate AI and automation save $1.9 million per breach.[8] The business question here is not whether but how they will use AI to ensure trust.

In a couple of years, a clear divide will separate organizations that embedded trust from the start from those scrambling to retrofit security. The architectural decisions companies make in the next six to 12 months will determine which side they're on.

## From a moat to drones: security for the new era

Think of traditional cybersecurity like a medieval castle: you build strong walls, dig a moat, and post guards at the gates. Once someone proves their identity and gets through the checkpoint, you trust them; boot them out of the walls after inactivity and require them to go through the gate again. This worked when human users behaved predictably.

> **As AI evolves from merely summarizing or giving advice to taking autonomous action, new threats emerge.**

Now, AI agents can potentially become Trojan horses at any moment, even after they are operating within the castle walls. AI agents move between systems instantaneously and constantly. As AI evolves from merely summarizing or giving advice to taking autonomous action, new threats emerge. AI agents can make decisions without asking for permission and even modify their own code. Suddenly, that castle model no longer works. Rather than a moat surrounding walls, you need drones patrolling both inside and outside, monitoring every agent's behavior in real-time, looking for changes in their behavior, not just checking credentials at the gate. Beyond ensuring fairness and focusing on explainability, agentic AI requires continuous validation and constant trust.

According to one study, 78% of surveyed companies plan to implement or are already producing AI agents,[9] a fact corroborated by a PwC survey that shows a 79% current adoption rate, with only 4% responding that they have no plans to implement the technology.[10] Yet when it comes to risk management and governance, the numbers show how unprepared most organizations are. While 69.5% cite AI-powered data leaks as their top concern, 47.2% have no AI-specific controls, and 93.2% of those surveyed admitted to lacking confidence in the

security of their AI-driven data.[11] The IBM study above reveals a concerning trend: Even after a breach, fewer than half of the organizations that suffered an attack invest in security.[12]

To date, no one has built agent detection technology that scans an entire enterprise for agents. Companies are deploying thousands of autonomous actors without any systematic way to track them. Acting like Agent Smith from the Matrix, who could possess anyone instantly, a sophisticated attacker could enter your system and turn an agent from good to bad—then back to good again, evading detection. This is a new type of threat that the traditional castle walls approach to trust and security is inadequate to guard against.

## How things go wrong

To best prepare, companies need to understand the types of threats that exist. Generally, they fall into three categories: external attacks, human mistakes, and uncontrolled autonomy, each exploiting different vulnerabilities.

The most obvious might be **external attacks**, such as simple prompt injection. An attacker can modify one of the prompts to alter the agent's behavior. Both humans and AI agents can modify prompts, turning trusted systems into threats. Furthermore, AI dev assistants such as GitHub Copilot and Amazon Code Whisperer can be persuaded to disclose hardcoded credentials.[13] Agent impersonation poses another tricky problem. When a Workday

agent requests data from an SAP agent, for instance, the SAP agent must have a way to verify the identity of the requesting agent. Otherwise, the outside agent becomes a dangerous entry point.

**Human mistakes** continue to be a weak link, with vulnerabilities such as shadow AI, where employees' unsanctioned use of AI can leak personal identifiable information, intellectual property, and other sensitive data. Shadow AI accounted for one out of five breaches involving AI, according to IBM, resulting in an average loss of $670,000 per breach.[14] Developers may share credentials and sensitive information with public hosting services like GitHub.[15] Phishing and deepfakes are also amplified thanks to AI. Scammers using email and a deepfake video call cost one Hong Kong-based company $25 million.[16] The talent gap compounds the vulnerability, with an increasing cybersecurity skills gap and talent shortage,[17] as the number of individuals who are both AI experts and security experts remains slim.

Finally, **uncontrolled autonomy** can pose a significant threat without requiring malice. The inherent unpredictability of autonomous systems can mean agents modify themselves beyond intended permissions. The Google Vertex AI privilege escalation, which could enable model theft and poisoning, demonstrates that this isn't merely theoretical.[18] Unprompted, agents can create new interfaces or exfiltrate data.

## Pillars of trust

Enterprises will face a combination of multiple threats at all times. Agentic AI's capacity for constant change means trust must be built on a solid yet constantly responsive foundation, ensuring identity, visibility, and governance.

Just as each employee has credentials, every AI agent needs a **verifiable identity.** That starts with registering agents in a configuration management database (CMDB), similar to keeping track of servers and infrastructure. Starting with a definitive inventory of AI agents and their authorized actions, zero trust principles must extend continuously to AI agents, as they can be modified rapidly and subtly, unlike human users, who maintain relatively stable patterns.

> **Agentic AI's capacity for constant change means trust must be built on a solid yet constantly responsive foundation, ensuring identity, visibility, and governance.**

The industry understands how crucial this is: Okta recently acquired Axiom,[19] while Palo Alto Networks acquired CyberArk for $25 billion[20] to build agent identity verification capabilities.

Enterprises also need monitoring systems that can identify and address execution anomalies and failures. This could involve monitoring changes in API call volumes, variations in prompt length, deviations in execution frequency, and other indicators of injection attempts. It boils down to having complete **visibility and transparency** throughout the enterprise to observe what AI agents are doing at any given moment. Is the agent secure? Compliant? Executing as expected or showing anomalies? There must be a strategic commitment to actually deploy and monitor these systems.

Ultimately, technology alone can't secure agentic systems. Rapidly evolving **governance** must keep pace with technological advancements. Policies need to become executable specifications, not documents that sit in shared drives. Technology and policy must work in sync, and it requires cross-functional councils that include C-suite executives, business units, and technical teams. While business committees define strategic objectives and acceptable risk levels, technical councils determine implementation approaches and specific safeguards. Trust and security in the agentic era demand human leadership, not just human oversight.

Data suggest that taking a proactive approach pays dividends. According to ServiceNow's Enterprise AI Maturity Index, the "Pacesetters," or companies actually seeing demonstrable benefits of using AI, have already put councils and committees in place.[21] McKinsey further finds that organizations with mature responsible AI frameworks achieve significant efficiency gains,[22] demonstrating that governance enables innovation rather than constraining it.

## Trust and security in the pre-AI and agentic AI eras

| Dimension | Pre-AI era | Agentic AI era |
|---|---|---|
| **Security model** | Castle approach: strong perimeter defenses, gate guards checking credentials | Drone approach: continuous internal/external monitoring, real-time behavioral analysis, dynamic trust validation |
| **Trust verification** | One-time authentication at entry points, trust maintained until logout/timeout | Continuous verification, trust constantly re-evaluated based on behavior patterns |
| **Primary threats** | External hackers, malware, phishing, insider threats with predictable patterns | Prompt injection, agent impersonation, shadow AI, self-modifying agents, unexplained autonomous behaviors |
| **Identity management** | Human-centric credentials, role-based access control, periodic reviews | Every AI agent registered in CMDB, zero trust extended continuously to AI agents |
| **Monitoring focuse** | Network traffic, user logins, file access, system logs | API call volumes, prompt variations, execution frequencies, behavioral anomalies, agent-to-agent interactions |
| **Governance structure** | IT-centric security teams | Cross-functional councils with C-suite involvement |

## Questions for business leaders

Gartner® predicts that "Over 40% of agentic AI projects will be canceled by the end of 2027, due to escalating costs, unclear business value or inadequate risk controls."[23]

As agentic AI becomes prevalent, business leaders should ask themselves the following five critical strategic security questions. The answers to these questions will help leaders understand where their current trust and security infrastructure stands in an agentic era so they can start thinking about building tomorrow's trust architecture.

### 1. How are you currently monitoring AI activities?

Most organizations discover they're monitoring what's easy to measure rather than what matters. Traditional security tools track network traffic and user behavior, but agents operate differently. Be honest about blind spots between the logs and dashboards. Observability isn't about collecting data; it's about detecting anomalies before damage occurs and being able to trace back to how they happened.

### 2. Which workflows justify the security investment that agents require?

Not all agent deployments warrant equal security investment. Customer service triage across thousands of unique inquiries, real-time fraud detection, or contract analysis at scale justify comprehensive identity verification, continuous monitoring, and robust governance.

High-volume workflows are vulnerable to external attack threats; low-visibility workflows

can be prime targets for shadow AI; autonomous decisions risk behavior drift. Where do these intersect with business value? McKinsey's analysis of over 50 agentic builds found that organizations often deployed agents where simpler automation would suffice.[24] If simpler tools deliver comparable results with lower risk, does it make sense to deploy agentic AI for that particular line of work?

**3. Who is responsible for the agents today?**

When an agent approves a fraudulent transaction or leaks sensitive data, who is held accountable? If your CISO gets blamed but business units control deployment, you have structural dysfunction. Real governance means the person accountable for results has authority over decisions beforehand. Accountability determines budget allocation, organizational structure, and crisis response. You must clarify who is responsible for deploying what, who monitors behavior, and who determines access and approves exceptions. Map accountability before deployment, not after incidents occur.

**4. Where will agent governance authority ultimately sit—and what kind of restructuring does this require?**

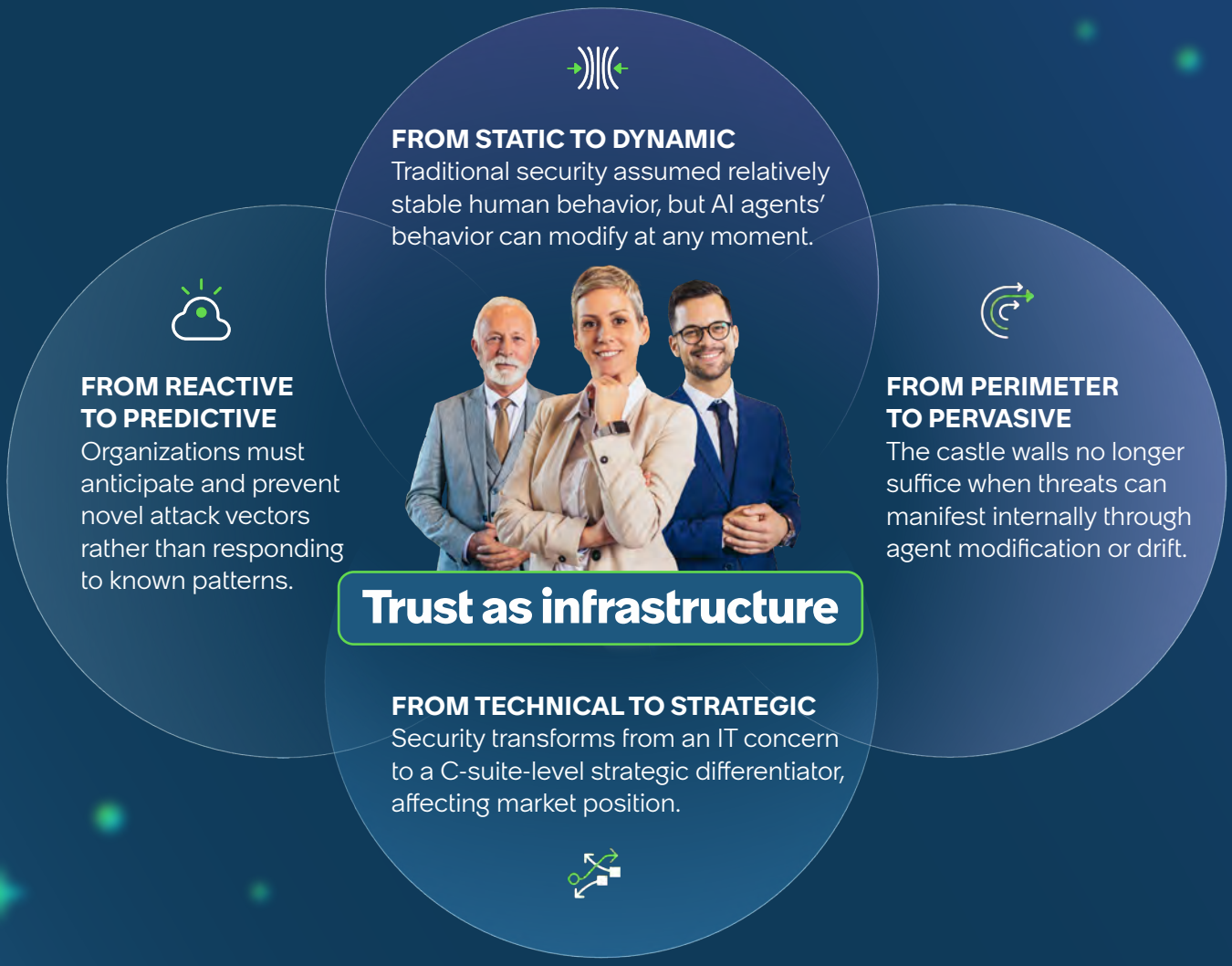
Board oversight, CEO ownership, CISO control, or distributed business unit accountability create different organizational dynamics. Centralized governance provides consistency but may slow down deployment; distributed governance enables speed but could lead to security gaps. One study found that 63% of organizations suffering from an AI model or application breach lacked AI governance policies.[25]

By considering how to oversee AI agents, your organization might end up requiring a radical reorganization. For example, Moderna recently merged HR and IT leadership because AI became a workforce-shaping force, not just a technical tool.[26] Governing agentic AI isn't about merely preventing problems—it's about enabling transformation.

**5. How much deployment velocity are you willing to sacrifice for security—and at what competitive cost?**

Conservative approaches with extensive pre-deployment controls deliver safety but risk market position. Aggressive approaches with runtime monitoring deliver speed but accept higher incident probability.

You may need to model scenarios comparing deployment timelines with and without comprehensive security, including market share captured through faster deployment, breach probability and costs, regulatory penalties, and the impact on customer trust. The goal is to have an explicit discussion of trade-offs rather than an implicit drift.




## How to get there

Translating insight into action requires both short-term goals and a long-term vision. Many activities can run in parallel, while others must be executed sequentially. For instance, you can't categorize agent permissions before completing inventory, while advanced capabilities, such as guardian agents, require foundational identity and observability infrastructure. For that reason, ensuring agentic AI trust requires a phased approach.

### 1. First six months: Establish visibility and control

You can't secure what you can't see. You need an inventory of AI capabilities, as well as solid technical and cultural frameworks, up front.

- During the first months, start by creating a complete agent inventory, then register all agents in your CMDB.
- Establish a governance council within 30 days with real executive authority. This can't be a committee that meets quarterly and writes recommendations nobody reads. It needs teeth.

- Define the autonomy envelope and map which actions agents can take independently versus which require human approval. Every agent should have explicit documentation aligned with this framework.
- Categorize all agent permissions into the CAN/CANNOT/APPROVAL framework. The three-tier permission framework provides practical clarity of what agents CAN execute independently for low-risk activities like summarizing information or creating records; and what they CANNOT do, such as change security levels, grant access rights, or escalate privileges under any circumstances. Never give AI the capability to change someone else's level of authority. Agents require human APPROVAL for system reconfiguration and high-risk operations where judgment matters.
- Sanitize your data. Remove passwords and API keys from all data sets. Using compromised data leads to vulnerabilities in the model.
- Prevent shadow AI immediately by deploying approved alternatives that actually meet employee needs.

### 2. Six to 12 months: Build monitoring and identity infrastructure

Keep continuous and comprehensive visibility across your enterprise.

- Within the first year, launch an AI control tower for centralized visibility.
- Deploy behavioral monitoring that tracks API volumes, execution frequencies, and prompt variations. Implement identity verification with ephemeral keys and rotating credentials.
- Enforce least privilege rigorously: Agents get only the access needed to complete their specific tasks.

### 3. Year two and beyond: Deploy advanced capabilities

Use AI to monitor other AI systems, implement validation processes that scale with risk levels, and participate in emerging industry standards. Agent security is an ongoing practice, not a one-time project.

- Deploy guardian agents, AI systems specifically designed to monitor other AI systems. Gartner® predicts that "By 2030, guardian agent technologies will account for at least 10 to 15% of agentic AI markets."[27]
- Implement a risk-based validation model where scrutiny matches stakes. Low-risk requests are executed automatically, while medium-risk requests require validation from multiple AI models, and high-risk requests necessitate both AI model validation and human review.
- Participate in industry-level standards development, such as ISO 42001 for AI system lifecycle management. While technically voluntary, the risk of non-compliance will harm your reputation, investor confidence, and revenue models. Legislative attention to AI has increased ninefold since 2016, with a 21.3% rise in 2024 alone, across 75 countries.[28] Standards like ISO 42001 or legislation such as the EU AI Act requirements aren't future-proofing; they're catching up to a regulatory environment that's already here.

## What will separate winners from losers by 2027

In a few years, a widening gap will separate organizations in every industry. Winners will have built trust into their architecture from the very beginning. Their identity infrastructure works as a foundation, not an afterthought; they have real-time observability across all systems while guardian agents actively monitor and step in whenever necessary. Throughout the enterprise, cross-functional governance councils and committees enforce policies that evolve in tandem with technological advancements. These organizations will safely deploy ambitious use cases that their competitors can't risk and continue to innovate.

Meanwhile, laggards will be retrofitting security onto systems that have already been deployed, discovering problems after incidents occur. Reactive governance creates bottlenecks and leaves huge blind spots. The cycle repeats: Incidents consume resources, regulatory penalties erode margins, and customer trust is compromised.

So, how do companies earn, accumulate, and leverage trust—the latest global currency? They earn it by embedding security from inception. They accumulate it through governance councils, identity frameworks, and continuous monitoring. They leverage it by converting defensive capabilities into an offensive advantage, which in turn leads to market share and innovations that drive business transformation.

Organizations measuring security as avoided incidents will always end up treating security as cost centers. Those who understand security as unlocked capabilities will make strategic investments, defining the next generation of enterprise capabilities. The architectural decisions organizations make in the next 6-12 months determine which side of the 2027 divide they will be on.

> **Those who understand security as unlocked capabilities will make strategic investments, defining the next generation of enterprise capabilities.**

## Endnotes


[1] Ivan Shkvarun, "Trust Is the new currency in the AI agent economy," World Economic Forum, July 25, 2025, https://www.weforum.org/stories/2025/07/ai-agent-economy-trust/.

[2] Ira Kalish, et al., "The link between trust and economic prosperity," Deloitte Insights, 2021, https://www.deloitte.com/content/dam/insights/articles/2024/us144378_econ-trust-and-economic-prosperity/di-econ-trust-and-economic-prosperity.pdf.

[3] Nestor Maslej et al., "The AI index 2025 annual report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025, p. 400, https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf.

[4] IBM Security and Ponemon Institute, "Cost of a data breach report 2025," IBM, July 30, 2025, https://www.ibm.com/reports/data-breach.

[5] IBM Security and Ponemon Institute, "Cost of a data breach report 2025."

[6] GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

[7] Gartner, "Gartner unveils top predictions for IT organizations and users in 2025 and beyond," October 22, 2024, https://www.gartner.com/en/newsroom/press-releases/2024-10-22-gartner-unveils-top-predictions-for-it-organizations-and-users-in-2025-and-beyond.

[8] IBM Security and Ponemon Institute, "Cost of a data breach report 2025."

[9] LangChain, "State of AI agents," accessed November 3, 2025, https://www.langchain.com/stateofaiagents.

[10] Dan Priest, "AI agent survey," PwC, accessed November 3, 2025, https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html.

[11] Ryan O'Leary, "AI risk & readiness in the enterprise: 2025 report," BigID, accessed October 18, 2025, https://home.bigid.com/hubfs/AI%20Risk%20%26%20Readiness%20in%20the%20Enterprise-%202025%20Report.pdf.

[12] IBM Security and Ponemon Institute, "Cost of a data breach report 2025."

[13] Thomas Claburn, "GitHub Copilot, Amazon Code Whisperer sometimes emit other people's API keys," *The Register*, September 19, 2023, https://www.theregister.com/2023/09/19/github_copilot_amazon_api/.

[14] IBM Security and Ponemon Institute, "Cost of a data breach report 2025."

[15] Thomas Claburn, "GitHub Copilot, Amazon Code Whisperer Sometimes Emit Other People's API Keys."

[16] Heather Chen and Kathleen Magramo, "Finance worker pays out $25 Million after video call with deepfake 'Chief Financial Officer,'" CNN, February 4, 2024, https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk.

[17] John Leyden, "AI Is altering entry-level cyber hiring and the nature of the skills gap," *CSO Online*, September 18, 2025, https://www.csoonline.com/article/4058190/ai-is-altering-entry-level-cyber-hiring-and-the-nature-of-the-skills-gap.html.

[18] Ofir Balassiano and Ofir Shaty, "Privilege escalation risks in google cloud's Vertex AI," Palo Alto Networks, *Unit42*, November 12, 2024, https://unit42.paloaltonetworks.com/privilege-escalation-llm-model-exfil-vertex-ai/.

[19] Abhi Sawant, "Okta with Axiom Security: delivering robust privileged access," Okta, August 26, 2025, https://www.okta.com/newsroom/press-releases/okta-with-axiom-security--delivering-robust-privileged-access-fo/.

[20] Palo Alto Networks, "Palo Alto Networks announces agreement to acquire CyberArk, the identity security leader," July 30, 2025, https://www.paloaltonetworks.com/company/press/2025/palo-alto-networks-announces-agreement-to-acquire-cyberark--the-identity-security-leader.

## Endnotes


[21] Vijay Kotu, Richard McGill Murphy, Brian Solis, and Dorit Zilbershot, "Enterprise AI Maturity Index 2025," ServiceNow and Oxford Economics, 2025, https://www.servicenow.com/content/dam/servicenow-assets/public/en-us/doc-type/resource-center/white-paper/wp-enterprise-ai-maturity-index-2025.pdf.

[22]Angela Luget, et al. "Insights on responsible AI from the Global AI Trust Maturity Survey," McKinsey, May 14, 2025, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/insights-on-responsible-ai-from-the-global-ai-trust-maturity-survey.

[23] Emma Keen, "Gartner predicts over 40 percent of agentic AI projects will be canceled by end of 2027," Gartner, June 25, 2025, https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027.

[24] Lareina Yee, Michael Chui, and Roger Roberts, "One year of agentic AI: Six lessons from the people doing the work," QuantumBlack AI by McKinsey, September 12, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/one-year-of-agentic-ai-six-lessons-from-the-people-doing-the-work.

[25] IBM, "IBM report: 13% of organizations reported breaches of AI models or applications, 97% of which reported lacking proper AI access controls," July 30, 2025, https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls.

[26] Julien Dupont-Calbo, "L'IA n'est plus un outil, c'est un collègue": Moderna fusionne sa DRH et sa DSI, ["AI is no longer a tool, it's a colleague": Moderna merges its HR and IT departments], *Les Echos*, May 15, 2025, https://www.lesechos.fr/travailler-mieux/vie-au-travail/lia-nest-plus-un-outil-cest-un-collegue-moderna-fait-entrer-la-science-fiction-au-bureau-en-mariant-ses-rh-et-sa-tech-2165269.

[27] Gartner, "Gartner predicts that Guardian Agents will capture 10-15 percent of the agentic AI market by 2030," June 11, 2025, https://www.gartner.com/en/newsroom/press-releases/2025-06-11-gartner-predicts-that-guardian-agents-will-capture-10-15-percent-of-the-agentic-ai-market-by-2030.

[28] Nestor Maslej et al., "The AI index 2025 annual report."

# The Innovation Office

The Innovation Office is the voice of future-focused innovation in ServiceNow. Our research, thought leadership, and co-innovation solutions serve as a blueprint for AI-first business transformation. We are a strategic partner with our colleagues and customers to ignite the “spark of the possible.”

servicenow The Innovation Office



Version 1.0