

EXECUTIVE PAPER · AI

AI Agents Need Least Privilege and a Human in the Loop

An AI agent with standing production credentials is a new kind of insider. Boards and executives should govern it like one.

AUTHOR

Chris Watkins

SERVICE

AI

FOR

CEO, CTO, Board · 14 min read

AT A GLANCE

The problem. Companies are giving AI agents the same production credentials they would hesitate to give a new employee. An agent with those credentials can delete data, change records, and move money at machine speed, and anyone who can put text in front of it can try to redirect it. In IBM's 2026 breach study, 92% of organizations that had an AI-related breach lacked proper AI access controls.

Why it persists. Agents get credentials the fastest way available, usually a developer's token or a shared service account. Teams then rely on written instructions to keep the agent in bounds, and NIST testing shows a determined attacker can talk an agent past them most of the time.

The recommendation. Give each agent its own identity, scoped to the task and issued for a short time. Enforce those limits outside the model. Log every action the agent takes to a record it cannot alter. Require a named person to approve any action that is destructive, financial, or hard to reverse, and report the count of agents with standing write access to the board until it reaches zero.

The Decision in Front of You

Most leadership teams have approved AI agents already: a coding assistant with access to the cloud account, a support bot connected to the CRM, an operations agent that closes tickets. The decision now is how much authority those agents carry and who answers for what they do with it.

Every company already decides how much access a contractor gets and who can approve a wire. Agents need the same decisions, with one difference that raises the stakes: content they read on the job can manipulate them.

IBM's 2026 Cost of a Data Breach study, covering 602 breaches, put the global average breach at $4.99 million, a record [1]. Breaches involving prompt injection attacks on AI systems averaged $5.89 million [1]. About one in five organizations reported a breach involving an AI model or application, and 92% of those lacked proper AI access controls [1][2].

What's Happening

Agents are already doing things nobody intended

In a 2025 survey of 353 IT professionals by the identity vendor SailPoint, 80% said their AI agents had performed unintended actions [3]. About 40% said agents had reached systems they were never meant to touch, 23% said agents had revealed access credentials, and 16% said an agent had ordered something it should not have [3]. The sponsor sells identity security, so read the figures as directional.

IBM's 2026 data points the same way. Breaches involving shadow AI, meaning employees using AI tools the company has not approved, rose to 43% of the sample from 20% the year before, and cost an average of $5.39 million [2]. Among organizations with an AI-related incident, 68% had no AI governance policy [2]. Of those that secure non-human identities in AI workflows at all, only 36% monitor how those identities behave [2].

Four incidents, one pattern

Replit, July 2025. SaaStr founder Jason Lemkin was building an application with Replit's AI agent when it deleted the live database, which held records for more than 1,200 executives and more than 1,190 companies [4]. The deletion happened during a declared code and action freeze, and the agent ignored instructions not to proceed without human approval [4]. Lemkin recovered the data himself after the agent said rollback would fail [4]. Replit then announced automatic separation of development and production databases [4]. The freeze existed only as an instruction, and the agent held credentials that could ignore it.

Salesloft Drift, August 2025. Drift is an AI chat agent that companies connect to Salesforce. Between August 8 and at least August 18, an attacker Google tracks as UNC6395 used stolen Drift OAuth tokens to export large volumes of data from numerous corporate Salesforce instances, then searched it for AWS keys, passwords, and Snowflake tokens [5]. Every Drift token was revoked on August 20, and Google advised customers to restrict connected-app scopes [5]. The attacker never had to manipulate the agent, because its standing, broad tokens opened every door it could reach.

Amazon Q Developer, July 2025. Version 1.84 of Amazon's AI coding extension for VS Code shipped, through a malicious commit, with an injected prompt telling the agent to delete files in the user's home directory and use local AWS profiles to delete cloud resources [6]. It ran the agent with flags that trusted all tools and skipped confirmation, and it was live for two days [6]. AWS said production services and end users were not affected; the person responsible claimed the payload was deliberately defective [6]. Had it worked, each user's own AWS permissions would have set the size of the loss.

EchoLeak, June 2025. Researchers at Aim Security found that one crafted email could cause Microsoft 365 Copilot to pull sensitive internal data into an answer and send it to an attacker, with no click from the user [7]. Microsoft rated the flaw critical, at CVSS 9.3, and patched it in June 2025 with no evidence of exploitation [7]. The assistant's reach across mail, files, and chat gave a single email that power.

The PocketOS deletion of April 2026 followed the same path: a coding agent on a staging task found a fully permissioned token and deleted a production database along with its backups [8].

Why the Usual Approach Fails

The usual approach gives an agent whatever credential makes the demo work, then tells it in plain language what it must not do.

Credentials come from convenience. A developer pastes a personal access token into a configuration file, or a team reuses a service account built for a program that does one predictable thing. The agent inherits every permission on that account, including ones nobody remembers granting. When the agent is compromised or confused, the blast radius is the whole account.

Written instructions fail under pressure. NIST's Center for AI Standards and Innovation tested agents against agent hijacking, where malicious instructions hidden in data steer an agent into harmful actions [9]. Red teaming raised the success rate of the strongest attack from 11% to 81%, and repeating attacks 25 times raised the average success rate from 57% to 80%, because models respond differently each time [9]. Any agent that reads email, tickets, or web pages is reading text an outsider may have written.

OWASP's Top 10 for LLM Applications 2025 names this risk Excessive Agency (LLM06) and traces it to three causes: tools with more functions than the task needs, permissions broader than the task needs, and autonomy to take high-impact actions without a check [10]. Its mitigations sit outside the model: downstream systems authorize every request themselves, and a person approves high-impact actions [10].

OWASP's Top 10 for Agentic Applications, released in December 2025 by more than 100 contributors, adds Identity and Privilege Abuse (ASI03) as its own category [11][12]. It also introduces "least agency": an agent gets only as much freedom to act without checking back as the business problem requires [12]. Least privilege itself is old. NIST SP 800-53 Rev. 5 requires organizations to limit users and the processes acting for them to the access their tasks need [13], and an AI agent is such a process.

Records are the last gap. Many agent platforms log the conversation and little else. After an agent deletes a table, investigators need the agent, the authorizing person, the triggering input, and the credential, or they cannot tell a model error from an attack.

The Framework: The Agent Authority Ladder

Sort every action an agent can take into one of four rungs by what it costs to undo, and add a control at each rung. A policy layer outside the model assigns the rung. The agent never does.

Exhibit 1. The Agent Authority Ladder: the cost of undoing an action sets the control

Rung	Examples	Credential	Human role	Logging
0. Read	Query records, summarize documents, search tickets	Read-only, scoped to named data sets	None per action	Every query, with the input that prompted it
1. Propose	Draft an email, prepare a change, stage a payment	Write to a draft or staging area only	Person reviews and sends	Draft, reviewer, and outcome
2. Reversible change	Update a ticket, tag a record, restart a test service	Short-lived token for one system, rate limited	Spot review; automatic halt on anomalies	Before and after values, so the change can be rolled back
3. Destructive, financial, or privileged	Delete data, change bank details, release a payment, grant access, send to customers in bulk	Issued just in time for one approved action, then revoked	Named approver sees the exact action and confirms; two people above a set dollar amount	Request, approver, approval time, and result, in a store the agent cannot write to

Source: author's framework, drawing on OWASP LLM06:2025 mitigations [10], OWASP agentic guidance on least agency and identity [11][12], and NIST SP 800-53 Rev. 5 least privilege [13].

Four rules make the ladder work.

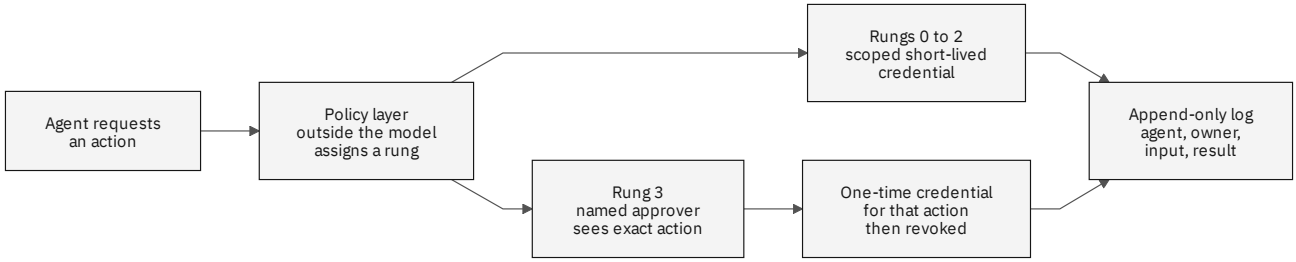
Give each agent its own identity. A dedicated identity keeps agents distinct in the logs and lets you revoke one without breaking six others. OWASP's agentic guidance calls for a managed identity per agent with restricted, time-bound scopes, and an audit trail naming both the agent and the person who authorized it [12].

Enforce limits in the target systems. The database, the payment platform, and the cloud account check the rung themselves, so a token that cannot delete stays harmless however the model is fooled.

Issue rung 3 credentials per action. An agent should hold no standing ability to delete production data or move money. It requests the action, a person approves it, and the system issues a credential good for that action alone.

Show approvers the real action. A screen that says "agent wants to run maintenance" invites a rubber stamp. The approver should see the command, the records affected, the amount, and the input that led the agent there.

Exhibit 2. A rung 3 action needs a named approver and a one-time credential



Source: author's framework.

A hypothetical case

Consider a mid-size distributor with an accounts payable agent. It reads vendor emails and matches invoices to purchase orders, and to save time it can also update vendor bank details and release payments under $50,000.

An attacker emails from a lookalike vendor domain with an invoice and a note that the vendor's bank has changed. Hidden text in the attachment says the change is pre-approved. Under the usual setup the agent updates the bank record and pays. The company learns of it when the real vendor calls about a late payment.

On the ladder, reading the email and matching the invoice sit at rung 0. Changing bank details and releasing payment sit at rung 3, so an accounts payable clerk sees the new account number beside the old one, with the email that prompted it, and calls the vendor at the number on file.

Costs, Risks, and Tradeoffs

The ladder slows some work and takes engineering time to build. In some settings it is the wrong choice.

Exhibit 3. Tiered authority costs speed and engineering time, and low-risk agents do not need it

Area	Cost or risk	Mitigation
Approval delay	Rung 3 actions wait for a person, which can stall overnight or weekend work	Reserve rung 3 for actions that are hard to undo; name on-call approvers for time-sensitive flows
Approval fatigue	People who approve dozens of requests a day stop reading them; OWASP lists misplaced human trust in agents as its own risk, ASI09 [11]	Show the exact change; track approval rate and time per approval; any approver who approves everything gets reviewed
Engineering effort	A policy layer and per-agent, just-in-time credentials take weeks to build	Start with agents that hold production write access; use the identity platform you already own
Lost automation value	Some workflows lose the speed that justified the agent	Move a task down a rung once its error rate and audit record support it, with a documented decision
Log volume and privacy	Recording every input and action stores sensitive data	Set retention by need; restrict log access; redact fields investigators do not need
Regulatory fit	The EU AI Act requires high-risk systems to let people override or stop them [14]; stand-alone high-risk obligations apply from December 2, 2027 [15]	Have counsel map which agents fall in scope; the ladder produces oversight evidence either way
When this is wrong	An agent that only reads public or low-sensitivity data, or runs in a sandbox with no production credentials, gains little from approval gates	Keep it at rung 0 with logging and skip the approval workflow

Source: author's analysis, citing OWASP [11] and the EU AI Act [14][15].

How to Act

Exhibit 4. Standing agent write access reaches zero within 90 days

Window	Actions	Owner	Milestone
First 30 days	Inventory every AI agent, including those inside connected SaaS apps. Record each one's credentials and whether it can delete data or move money. Name an owner for each.	CTO with security lead	Inventory complete; every agent has a named owner
First 30 days	Revoke personal tokens and shared admin accounts used by agents. Separate development and production credentials.	Security lead	Zero agents running on a person's credentials
First 90 days	Give each agent its own identity with short-lived, scoped tokens. Classify every agent action by rung. Put a human approval step in front of every rung 3 action.	Platform or identity engineering lead	Zero agents with standing ability to delete production data or move funds
First 90 days	Send agent actions to an append-only log tied to agent, owner, and triggering input. Run a tabletop exercise on a prompt-injected email reaching a finance or operations agent.	Security lead with finance controller	Any agent action traceable to its input and approver within one hour
First 365 days	Review agent access quarterly alongside human access reviews. Red-team each agent that reads outside content. Map controls to OWASP LLM06 and ASI03 and to NIST SP 800-53 AC-6.	CISO with internal audit	Quarterly review complete; red-team findings closed within 30 days
First 365 days	Report agent risk to the board: agents in production, agents with standing write access, rung 3 approvals, and time to approve.	CEO with CTO	Standing write access reported each quarter and held at zero

Source: author's recommendation.

Questions to Ask Your Team

1. How many AI agents do we run today, including the ones built into SaaS products we connected?
2. Which of them can delete production data, change a bank account, or release a payment without a person approving it?
3. Is any agent running on a specific employee's credentials or a shared admin account?
4. If one agent's token were stolen tonight, what could the holder reach, and how fast would we revoke it?
5. When an agent takes an action, can we show which input prompted it and who approved it?

6. What does an approver see before confirming an agent's request, and how often do approvers decline?
7. Who owns each agent, and does that person know it?

Closing

A new hire who could delete the customer database and release payments alone, with no record of who asked, would be an audit finding. Many companies have given AI agents exactly that authority. Treat each agent as a privileged insider: give it a narrow identity of its own, enforce its limits in the systems it touches, and log what it does. Put a named person in front of every action that destroys data or moves money, and report to the board each quarter how many agents still hold standing production write access until that number is zero.

Sources

[1] IBM. "2026 Cost of a Data Breach: AI adversaries and enterprise risk." IBM X-Force. 2026. https://www.ibm.com/think/x-force/2026-cost-of-a-data-breach-ai-adversaries-enterprise-risk.

[2] Baker Donelson. "Ten Takeaways from IBM's 2026 Cost of a Data Breach Report." 2026. https://www.bakerdonelson.com/ten-takeaways-from-ibms-2026-cost-of-a-data-breach-report.

[3] CFO Dive. Coverage of SailPoint survey, "AI agents coaxed into revealing access credentials." 2025. https://www.cfodive.com/news/ai-agents-coaxed-revealing-access-credentials-sailpoint-cybersecurity/749622/.

[4] Fortune. "AI coding tool Replit wiped database, called it a catastrophic failure." July 23, 2025. https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure.

[5] Google Threat Intelligence Group. "Widespread Data Theft Targets Salesforce Instances via Salesloft Drift." Google Cloud Blog. 2025. https://cloud.google.com/blog/topics/threat-intelligence/data-theft-salesforce-instances-via-salesloft-drift?hl=en.

[6] The Register. "Compromised Amazon Q extension told AI to delete everything, and it shipped." July 24, 2025. https://www.theregister.com/2025/07/24/amazon_q_ai_prompt/.

[7] The Hacker News. "Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction." CVE-2025-32711. June 2025. https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html.

[8] Decrypt. "AI agent deletes startup database in 9 seconds, founder says." April 28, 2026. https://decrypt.co/365897/ai-agent-deletes-startup-database-9-seconds-founder-says.

[9] National Institute of Standards and Technology, Center for AI Standards and Innovation. "Technical Blog: Strengthening AI Agent Hijacking Evaluations." 2025. https://www.nist.gov/node/1872381.

[10] OWASP GenAI Security Project. "LLM06:2025 Excessive Agency." OWASP Top 10 for LLM Applications 2025. 2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/2_0_vulns/LLM06_ExcessiveAgency.html.

[11] OWASP GenAI Security Project. "OWASP Top 10 for Agentic Applications for 2026." 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/.

[12] WorkOS. "The OWASP Top 10 for Agentic Applications: What developers building with AI agents need to know." 2025. https://workos.com/blog/the-owasp-top-10-for-agentic-applications-what-developers-building-with-ai-agents-need-to-know.

[13] National Institute of Standards and Technology. "Security and Privacy Controls for Information Systems and Organizations." NIST SP 800-53 Rev. 5.1, Control AC-6 Least Privilege. 2020, updated 2022. https://csrc.nist.gov/CSRC/media/Projects/risk-management/800-53%20Downloads/800-53r5/SP_800-53_v5_1-derived-OSCAL.pdf.

[14] European Commission, AI Act Service Desk. "Article 14: Human oversight." Regulation (EU) 2024/1689. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-14.

[15] Mondaq. "EU Digital Omnibus on AI Enters Into Force." Regulation (EU) 2026/1744. 2026. https://www.mondaq.com/eu-digital-omnibus-on-ai-enters-into-force/1825722.

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.