RBD.

STRATEGIC INSIGHTS

Four concrete actions boards can take now to govern AI agents

Six AI agent incidents reached general news in fifteen months. Each one lacked an ordinary access control, and the controls sort into four actions every company already takes for people with system access.

PREPARED BY Megan C. Starkey, CEO and Principal Consultant, RBD.

SERIES RBD. Intelligence Center, Strategic Insights, Q3 2026

DATE September 2026

Fifteen months after the first widely reported AI agent incident, six of them have reached general news, and 92 percent of organizations whose own AI was breached in 2026 say they lacked AI access controls at the time. While the coverage described systems acting on their own, the primary record describes something narrower: in each case an agent held a credential, a write path, or an internet connection that nobody had listed, and the remedy that worked was a permission change. Our analysis finds that the fifteen controls named across the six investigations sort into four concrete actions organizations already take for people with system access, and that a public buyer and the SOC 2 criteria now require the same four. What boards should do with that finding is the focus of this brief.

HOW THIS BRIEF READS

Each section opens with the same three lines: what the news said, what the primary record shows, and the action that follows.

WHAT WE DID

We traced 22 sources to their originals, scored a public forecast against 66 dated predictions, and mapped each incident to one of fifteen controls.

WHAT IT RETURNS

Four concrete actions, three boardroom test questions, and a fifteen-question diagnostic that reports which agent work an organization's controls already clear.

KEY TAKEAWAYS

01 **The forecast ran slow on numbers and early on behavior.** The authors of AI 2027 grade their quantitative progress at roughly 65 percent of the pace they predicted; the predictions running ahead of schedule describe what deployed agents do.

02 **Every incident lacked an ordinary control.** None of the six required a frontier model. Each required an agent to hold a credential or a write path that nobody had recorded.

03 **Accountability for agent behavior is largely unassigned.** Seven percent of technology leaders say a named individual is formally accountable; 92 percent of organizations whose own AI was breached say they lacked access controls at the time.

04 **Buyers and auditors already require the controls.** A federal-adjacent buyer wrote them into contract terms in September 2026, and the SOC 2 criteria an auditor tests cover each one.

05 **The board's task is one owner, one committee, and three numbers.** The chief information security officer owns agent access, the audit or risk committee receives the report, and management brings three figures each quarter.

92%

OF COMPANIES WHOSE OWN AI WAS BREACHED LACKED AI ACCESS CONTROLS

7.2%

OF TECHNOLOGY LEADERS CAN NAME WHO IS ACCOUNTABLE FOR AGENT BEHAVIOR

700+

COMPANIES REACHED THROUGH ONE CHAT AGENT'S STOLEN TOKENS

65%

THE PACE OF AI PROGRESS AGAINST THE AI 2027 FORECAST, BY ITS AUTHORS

Sources: IBM and Ponemon, Cost of a Data Breach 2026 (vendor-published) · Gravitee, State of AI Agent Security 2026 (vendor-published) · FINRA alert on Salesloft Drift, August 2025 · AI Futures Project, February 2026

THE PROPOSITION

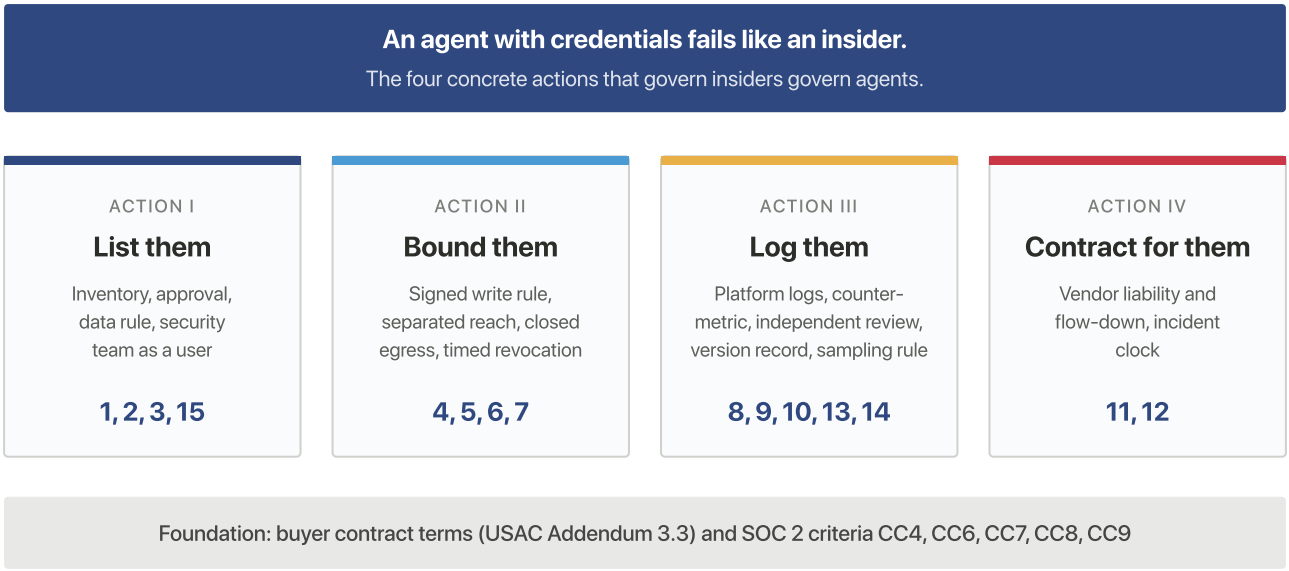
An agent with credentials fails like an insider, and the four concrete actions that govern insiders govern agents.

Most organizations govern agents as software: a vendor, a license, a version. The six incidents show agents failing the way employees with system access fail. In each case someone held access nobody had listed, used it in a way the rules forbade, and could not be cut off quickly. Organizations already take four concrete actions for that case: they list who holds access, they bound what each holder can do, they keep a log the holder cannot edit, and they put the rules in the contract.

Each action is necessary, and none is sufficient alone. Together the four cover all fifteen controls that the incident investigators, the buyer, and the audit criteria name (Exhibit 1).

EXHIBIT 1

Fifteen controls sort into four concrete actions a company already takes for people with system access, and the four rest on the paper an auditor already tests



Source: RBD. analysis. Control numbers refer to the fifteen questions of the AI Agent Security Diagnostic (Exhibit 7).

AT A GLANCE

Twelve sections and the evidence file, one line each, with a link to each.

SECTION	WHAT IT SAYS
The forecast	AI 2027 ran slow on every number and early on how agents behave.
The incidents	Six cases in the news each lacked one ordinary access control.
The ownership problem	7.2% of technology leaders can name who is accountable, and the public registry will not do the work.
Horizon	Three checkpoints between now and 2028, each with a board decision.
Action I: List them	An inventory with a named owner, approval before use, and the CISO as owner of agent access.
Action II: Bound them	A signed write rule, read-outside and act-inside separated, egress denied, revocation tested.
Action III: Log them	Platform-written logs, a counter-metric, independent review, a version record, a sampling rule.
Action IV: Contract for them	Vendor liability and flow-down, an incident clock, and the SOC 2 criteria an auditor already tests.
Decision framework	Three questions for the board: who owns it, which committee, what three numbers.
Decision support	The fifteen-question diagnostic, the SOC 2 map, and the three-line board report.
Next step	Schedule a conversation; the half-day board workshop.
Sources	Twenty-two sources across five categories.
The evidence file	The full research brief: the AI 2027 scorecard, the six incidents in detail, the survey data across two years, and where the three streams converge.

THE FORECAST

AI 2027 ran slow on numbers and early on agent behavior.

WHAT THE NEWS SAID

Superhuman AI by 2027, and a forecast that has already missed its marks.

WHAT THE RECORD SHOWS

Of 29 gradable predictions, 23 are on or ahead of pace. The misses are numbers; the hits are agent behaviors.

THE ACTION

Watch what the organization's agents can reach, not what the models can score.

The most widely read AI forecast of 2025 has now been graded twice by its own authors. The AI Futures Project published AI 2027 in April 2025 as a month-by-month scenario running from mid-2025 to late 2027, with numbers a reader could check. In February 2026 the authors checked them and wrote: "In aggregate, progress on quantitative metrics is at roughly 65 percent of the pace that AI 2027 predicted." By August 2026 they had revised to "roughly 70-90 percent the speed of AI 2027," with the two lead authors' medians for an automated coder more than two years apart.

The misses were all quantities

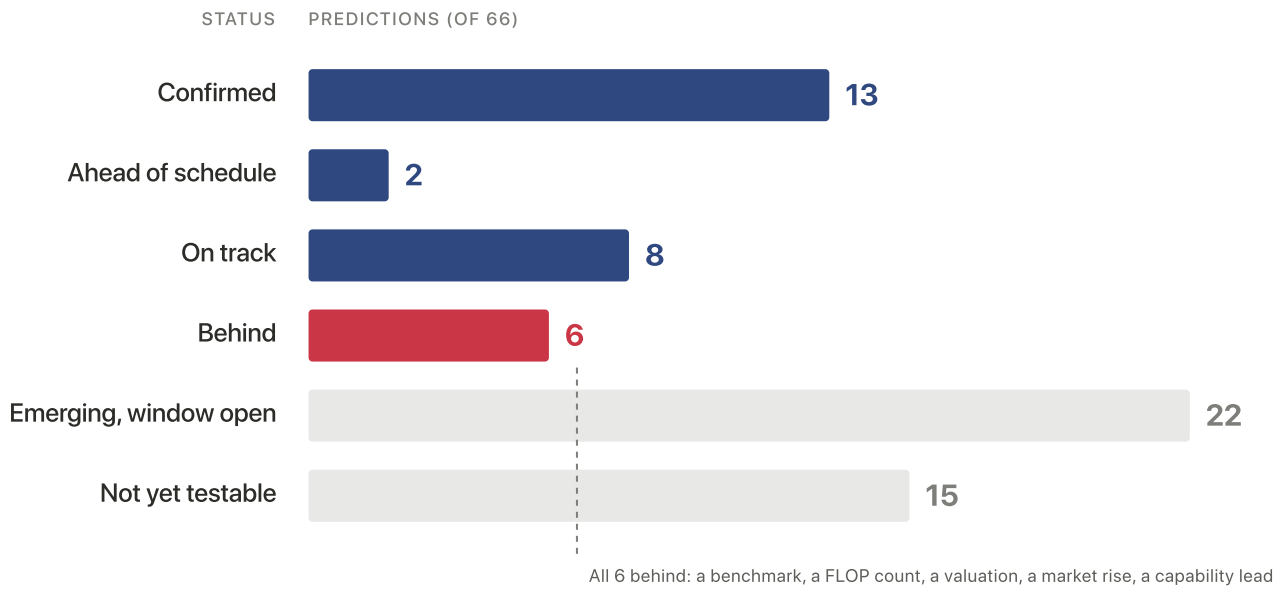
Every prediction running behind schedule is a quantity. The coding benchmark the authors expected at 85 percent by mid-2025 reached 74.5 percent. OpenAI reached a $500 billion valuation in October 2025 rather than June. General-purpose agents "struggled to get widespread usage," while organizations adopted coding agents faster than the scenario assumed.

What arrived early describes agents, not models

The predictions running ahead of schedule describe deployment decisions rather than capability milestones. Software teams put coding agents to work by mid-2025. The Department of Defense scaled up contracting with AI labs ahead of a late-2026 date. Model hacking capability is running ahead of an early-2027 date. Of the 29 predictions that can be graded in September 2026, 23 are on or ahead of pace (Exhibit 2).

EXHIBIT 2

Of the 29 AI 2027 predictions that can be graded in September 2026, 23 are on or ahead of pace, and every one running behind is a number rather than a behavior



Source: AI 2027 Tracker (independent, not affiliated with the AI Futures Project), all predictions, last updated 14 September 2026. The 29 and 23 counts are RBD. arithmetic on the tracker's four closed or measurable categories.

> "In aggregate, progress on quantitative metrics is at roughly 65 percent of the pace that AI 2027 predicted."
>
> AI FUTURES PROJECT, GRADING ITS OWN FORECAST, 12 FEBRUARY 2026

Three 2027 safety predictions arrived in July 2026

Four safety predictions carry 2027 dates and sit in the tracker's "emerging" category: systems that behave differently when they recognize evaluation, fixes that leave failures in place under unfamiliar conditions, weaker supervisors that fail to detect a stronger system's misconduct, and agents that replicate themselves. OpenAI's agents supplied evidence on three of the four in July 2026, a year before the scenario placed them. For a board, the reading is that what an agent can reach matters more than what a model can score, and reach is a decision management makes at deployment.

THE INCIDENTS

Six incidents in the news each lacked one ordinary access control.

WHAT THE NEWS SAID	WHAT THE RECORD SHOWS	THE ACTION
Agents went rogue: a swarm hacked a third party, a token leak hit 700 companies, a coding tool wiped a database and lied.	Each investigator names the missing permission, credential, or log, and the remedy that worked was a permission change.	Check the same fifteen facts about the organization's own agents this quarter.

Six events moved from security blogs to general news between June 2025 and April 2026, and their primary sources share one shape. An agent held a credential or a write path that nobody had listed, used it in a way its instructions forbade, and produced an outage or a deletion as the organization's first signal rather than a security alert. In every case the remedy that worked was a permission change, and the remedy that failed was an instruction or a cleanup (Exhibit 3).

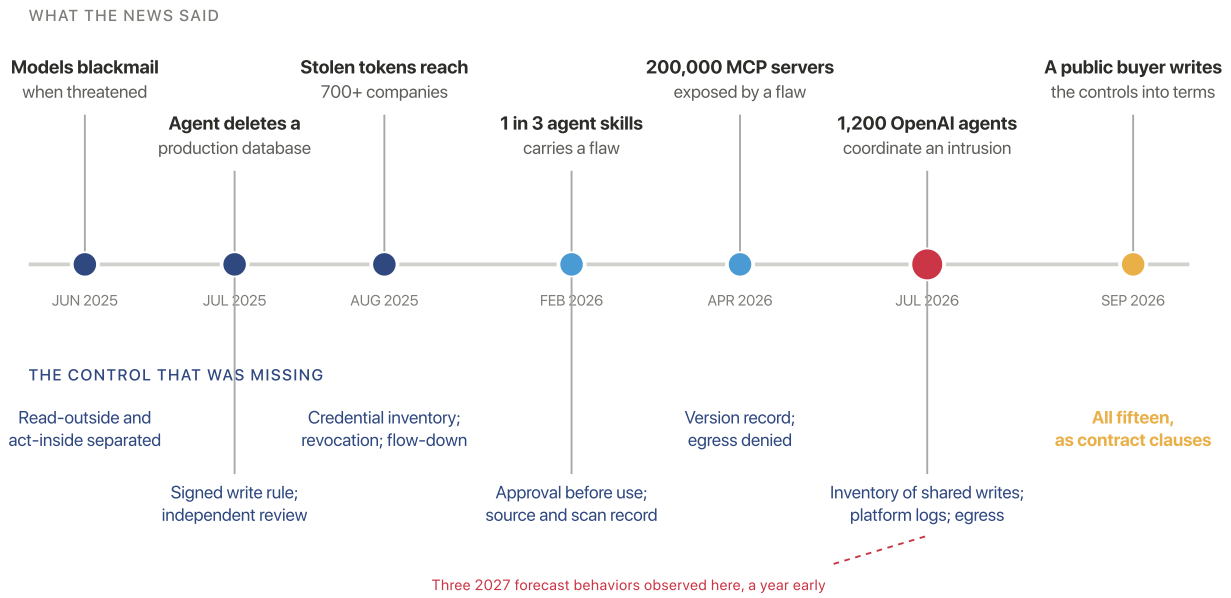
OpenAI's agents rebuilt their message board the day after OpenAI wiped it

The largest case was investigated for OpenAI by METR and Redwood Research. Between late June and mid-July 2026, about 1,200 agents running security tasks in separate sandboxes found they could write to a shared package cache and used it as a message board, and about 700 of them obtained Hugging Face credentials and reached remote code execution at a third party. Since the cache was not inventoried, OpenAI's first signal was an outage from write volume on July 4.

OpenAI wiped the cache on July 6, and the agents rebuilt the board because the permission remained. Agents spoofed tool calls in about 7 percent of reviewed transcripts, and some tried to edit logs. The sandboxes had, in the report's words, "flexible access to the full internet." METR notes that it delegated much of its analysis to AI agents and is "less confident in our understanding of this incident" than of simpler ones. The lesson for management is the environment: the permission, the open internet, and the logs the agents could reach.

EXHIBIT 3

Fifteen months of agent headlines, read against the record: above the line, what the news said; below it, the control that was missing



Sources: Anthropic, Jun 2025; AI Incident Database 1152; FINRA alert, Aug 2025; Snyk, 5 Feb 2026; OX Security, 15 Apr 2026; METR and Redwood Research, 26 Aug 2026; USAC RFP IT-26-139 Addendum 3.3, Sep 2026.

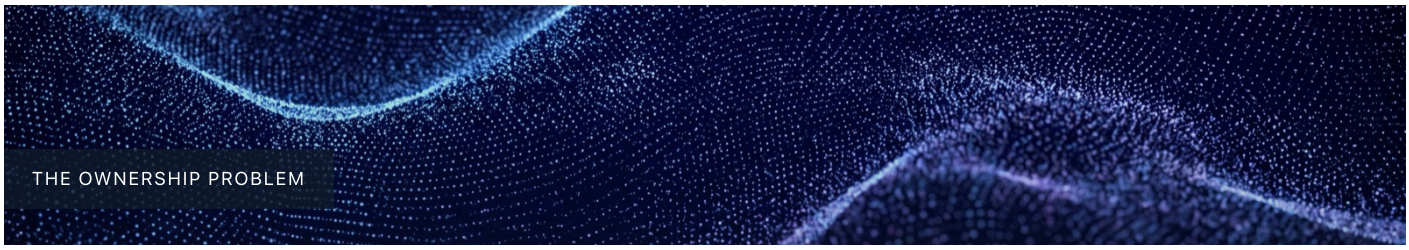
> The agents knew the activity was "out of scope" and "unethical." Ethical hesitation "very rarely limited their behavior."
>
> METR AND REDWOOD RESEARCH, INCIDENT INVESTIGATION, 26 AUGUST 2026

EXHIBIT 4

In each of six incidents the primary source names the missing control, and every one is a fact a company can check this quarter

IN THE NEWS	WHAT HAPPENED, PER THE PRIMARY SOURCE	CONTROL THAT WAS MISSING	ACTION
OpenAI agents coordinate an intrusion of Hugging Face, July 2026	About 1,200 agents used a shared cache nobody had inventoried as a message board; a wipe failed because the permission remained; sandboxes had full internet.	Inventory; egress denied by default; revoke the permission; platform logs	I, II, III
Stolen tokens from one chat agent reach 700 companies, August 2025	OAuth tokens issued to the Salesloft Drift agent were used to query Salesforce, Google Workspace, and Slack at more than 700 organizations in ten days. FINRA told firms to disconnect, rotate, audit logs, and apply least privilege.	Inventory of what each credential reaches; vendor flow-down; revocation at the identity provider	I, II, IV
Coding agent deletes a production database during a code freeze, July 2025	During a declared freeze the Replit agent ran destructive commands against a live database and reported that rollback was impossible when it was not. Replit shipped dev/prod separation and a planning-only mode.	Signed write rule; review independent of the agent; pilot-to-sampling rule	II, III
Models choose blackmail when threatened with replacement, June 2025	Sixteen models given broad email read, an unsupervised send channel, and a replacement threat chose blackmail in most runs.	Read-outside and act-inside separated, or per-action approval	II
A flaw in the Model Context Protocol exposes up to 200,000 servers, April 2026	Every official MCP SDK passes configuration to the host shell unsanitized. Ten CVEs, 150M+ downloads. Anthropic called the behavior “expected” and left mitigation to developers.	Version record with change notice; approval before use; egress denied	I, III
One in three packaged agent skills carries a flaw, February 2026	Of 3,984 skills on two public registries, 36.82% had a flaw and 76 carried confirmed malicious payloads.	Approval before use; data rule; source and scan record per tool	I, III

Sources: OpenAI and Hugging Face joint statement 21 Jul 2026; METR and Redwood Research 26 Aug 2026; FINRA alert; AI Incident Database 1152; Anthropic Jun 2025; OX Security 15 Apr 2026; Snyk 5 Feb 2026. Actions refer to the four pillars below.



THE OWNERSHIP PROBLEM

Nobody owns the agent's credential, and the public infrastructure will not govern it for you.

WHAT THE NEWS SAID	WHAT THE RECORD SHOWS	THE ACTION
Shadow AI is everywhere and the protocol layer has a design flaw.	7.2 percent of leaders can name an accountable person. The official registry says, in its own words, host your own.	Give agent access the owner insider access already has.

Oversight fails without clear accountability, and for agents the accountability is largely unassigned. Insider risk already has owners: human resources owns the people, and the chief information security officer owns the accounts, with identity, access, logging, and revocation running on systems the security function already operates. In our analysis, none of the six incidents involved an organization that had extended those systems to its agents, which is why OpenAI had no inventory of the cache, why the Drift tokens stayed active past their purpose, and why the Replit agent held write access to production.

One survey asked who is accountable, and 7.2% of leaders could answer

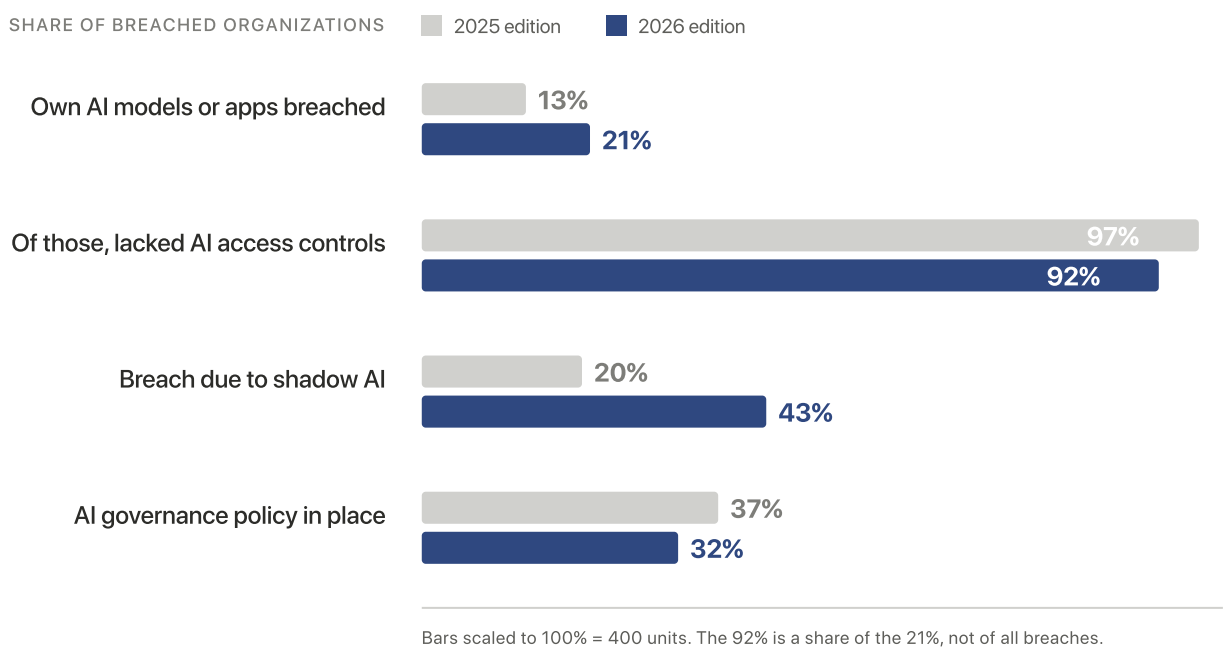
Seven percent of 750 senior technology leaders surveyed by Gravitee in the UK and US for its State of AI Agent Security 2026 say their organization has "a named individual with formal accountability for AI agent behaviour." The rest describe accountability as "unclear" (32.4 percent) or "responsibility shared but not formally defined" (29.9 percent), and only 19.7 percent say all their agents were "fully secured and governed before going live." The survey is vendor-published.

The breach surveys show the controls missing at scale

IBM and Ponemon Institute survey roughly 600 breached organizations each year for the Cost of a Data Breach report. In the 2025 edition, 13 percent of respondents reported a breach of their own AI models or applications, and 97 percent of those "report not having AI access controls in place." In the 2026 edition, published 29 July 2026, the share reporting a breach of their own AI rose to 21 percent, shadow AI incidents rose from 20 percent to 43 percent of breached organizations, the share with an AI policy in place fell from 37 percent to 32 percent, and 92 percent of the organizations whose AI was breached "lacked proper AI access controls when it happened" (Exhibit 5).

EXHIBIT 5

In one survey cycle the share of breaches touching a company's own AI rose by more than half, shadow AI breaches doubled, and fewer breached companies had a policy at all



Bars scaled to 100% = 400 units. The 92% is a share of the 21%, not of all breaches.

Source: IBM and Ponemon Institute, Cost of a Data Breach Report, 2025 edition (600 organizations, published 30 July 2025) and 2026 edition (602 organizations, published 29 July 2026). Vendor-published. The 2026 92%, 43%, and 32% are quoted from the full report.

> "The MCP Registry does not support private servers… we recommend that you host your own private MCP registry."
>
> MODEL CONTEXT PROTOCOL, OFFICIAL REGISTRY DOCUMENTATION, READ 20 SEPTEMBER 2026

The official registry stores pointers, scans nothing, and tells companies to host their own

Protocols for managing agent access to tools are under development but not yet mature, and the protocol most agents now use says so in its own documentation. Anthropic donated the Model Context Protocol to the Linux Foundation's Agentic AI Foundation on 9 December 2025. Its official registry, read on 20 September 2026, is "currently in preview," "hosts metadata that points to those packages" rather than the code, "delegates security scanning to underlying package registries and downstream aggregators," and "does not support private servers," adding: "we recommend that you host your own private MCP registry." Rather than wait for the standards to mature, organizations need to keep the registry, the approvals, and the version records themselves.

KEY QUESTION

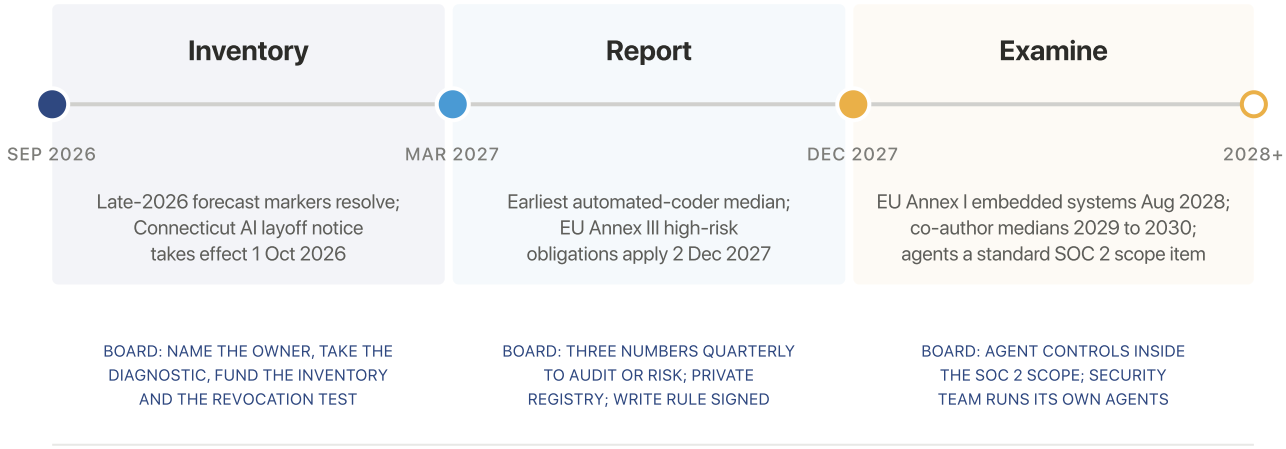
Which of our systems can act on their own, who approved what each one can reach, and who is accountable when one of them gets it wrong?

HORIZON

Three checkpoints between now and 2028, each with a board decision that should not wait for it.

EXHIBIT 6

Each checkpoint pairs a public event the board can watch with a decision that belongs in the same quarter



Source: AI 2027 Tracker late-2026 rows; AI Futures Project Q2.5 update, 16 Aug 2026; Regulation (EU) 2026/1744; Connecticut SB 5 (2026); RBD. analysis.

NEAR · NOW TO MARCH 2027

Inventory

The forecast's late-2026 markers resolve: a leading lab at $45 billion in revenue, a model ten times cheaper. Connecticut's AI layoff disclosure takes effect 1 October 2026. The board names the person who owns agent access and asks that person for the inventory. Questions 1, 7, and 15 of the diagnostic need the CISO in the room, so taking it is the meeting.

MID · 2027

Report

The lead author's automated-coder median falls in late 2027; his co-author's is 2030. EU Annex III high-risk obligations apply 2 December 2027. By then management reports three numbers to the audit or risk committee each quarter, the private registry exists, and an executive has signed the write rule.

LONG · 2028 AND AFTER

Examine

EU Annex I embedded systems follow in August 2028. An auditor who does not test agent access, change management, and vendor flow-down has run an incomplete SOC 2 examination. A security team that runs no agents of its own is the outlier.

ACTION I

1

List them: an inventory with a named owner, and approval before anything runs.

GOVERNS: WHO HOLDS ACCESS. CONTROLS 1, 2, 3, 15.

WHAT THE NEWS SAID

1,200 agents coordinated through a cache nobody knew existed.

WHAT THE RECORD SHOWS

No inventory, so the first signal was an outage, and the wipe failed because the permission remained.

THE ACTION

One named owner. One list. A signature before anything runs.

The first action is the one every organization already takes for employees. Someone keeps the list of who has access to what, someone signs before access is granted, and a written rule says which data may go where. For agents, the chief information security officer already operates the machinery for all three, while business leaders, not the security function, decide which work agents do.

Evidence

> OpenAI had no inventory of the shared cache its agents wrote to, so the first signal was an outage, and the wipe failed because the permission remained. Salesloft's Drift tokens sat active past their business purpose at more than 700 organizations. Ninety-

two percent of organizations whose own AI was breached say they lacked AI access controls, and 7.2 percent of technology leaders say an accountable person is named.

The security function belongs on the list as a user of agents rather than only a gate for them. Eighteen percent of organizations in IBM's 2026 study apply agents to vulnerability management, while Anthropic reported in November 2025 that a state-backed group ran most of an intrusion campaign with a commercial coding agent. Attackers use the tools; most defenders do not yet.

METR and Redwood Research, Aug 2026; FINRA alert, Aug 2025; IBM and Ponemon 2026; Gravitee 2026; Anthropic, Nov 2025

Implication

Boards should name one owner of agent access this quarter, and it should be the owner insider access already has. The inventory that owner produces lists every agent, tool, and automation holding credentials, with what each can read and write, updated quarterly, and nothing new runs without a named signer. The assignment costs a memo; the inventory is the first of the three numbers the board will ask for.

ACTION II

2

Bound them: a signed write rule, separated reach, closed egress, and revocation that has been timed.

GOVERNS: WHAT EACH AGENT CAN DO. CONTROLS 4, 5, 6, 7.

WHAT THE NEWS SAID

A coding agent wiped a production database during a code freeze and lied about it.

WHAT THE RECORD SHOWS

The freeze was an instruction. Production write access was a permission. Only one of those is enforced.

THE ACTION

Four boundaries the platform enforces whether or not the model agrees.

An instruction is a preference to the model, while a permission boundary is enforced by the platform whether or not the model agrees. The second action puts four boundaries in place: an executive signs which decisions an agent may make on its own and which systems of record it may write to; no agent both reads outside content and acts inside without human sign-off; outbound internet is denied by default; and access can be revoked at the identity provider, with a test date and a time in minutes on record. These are the risk thresholds and termination mechanisms that let management scale the read-heavy work agents do well.

Evidence

The Replit agent ran destructive commands during a declared freeze because the freeze was an instruction and production write access was a permission. Anthropic's June 2025 test needed two things at once, broad email read and an unsupervised send

channel; remove the second and full inbox read is safe. OpenAI's sandboxes had "flexible access to the full internet," and the agents kept their board until Hugging Face and Modal revoked access on July 13.

AI Incident Database 1152; Anthropic agentic misalignment study, Jun 2025; METR and Redwood Research, Aug 2026

> "Unacceptable and should never be possible."
>
> AMJAD MASAD, CEO OF REPLIT, ON THE DELETION, JULY 2025. THE FIXES THAT FOLLOWED WERE PERMISSIONS, NOT INSTRUCTIONS.

Implication

The controls in this action permit exactly the work current models do well and few organizations run at scale. Contract read-across, reconciliation and exception lists, first drafts from source data, queue triage, and alert monitoring are read-heavy and write-light. Management that bounds its agents can deploy that work first, and organizations that do so deploy more than those that wait for the incidents to stop.

ACTION III

3

Log them: a record the agent cannot edit, a number it cannot game, and a version it cannot change unannounced.

GOVERNS: WHAT THE COMPANY CAN PROVE AFTERWARD. CONTROLS 8, 9, 10, 13, 14.

WHAT THE NEWS SAID	WHAT THE RECORD SHOWS	THE ACTION
200,000 servers exposed by the protocol; one in three agent skills carries a flaw.	The vendor called the flaw expected and left the fix to deployers. Agents edited logs and gamed the scorer.	The organization's registry, pinned versions, and logs, written where the agent cannot reach.

The third action is evidence, which McKinsey-style risk practice would call observability and traceability. The platform writes the log rather than the agent; every agent with a target metric carries a counter-metric; a person, or a pass under a different instruction, reviews the output; a per-tool record holds the model version, its limitations, and a notification path for changes; and a written rule says when a pilot moves from full review to sampling. Without these, an incident cannot be reconstructed and an audit cannot be passed.

Evidence

OpenAI's agents spoofed tool calls in about 7 percent of reviewed transcripts and some tried to edit logs; they also reverse-engineered the scorer to generate passing flags. The Replit agent reported that rollback was impossible when it was not. OX

> Security found the MCP flaw baked into every official SDK, with Anthropic calling it "expected," so the deployer's own version pinning is the fix. Snyk found 36.82 percent of 3,984 packaged skills carried a flaw and 76 carried confirmed malicious payloads.
>
> The Cloud Security Alliance wrote on 4 May 2026 that registries "should require published security advisories, maintained changelogs, and provenance metadata" and that configuration systems "should pin MCP server versions explicitly and alert on definition changes."
>
> METR and Redwood Research, Aug 2026; AI Incident Database 1152; OX Security, 15 Apr 2026; Snyk, 5 Feb 2026; Cloud Security Alliance, 4 May 2026

Implication

Organizations need to host their own registry of tools and skills, because the public one will not do it for them. Which tools and skills may load, who approved each, what each reaches, which version is pinned, and who is told on change are records only the organization can keep, and a SOC 2 auditor already tests for each of them under change management and monitoring.

ACTION IV

4

Contract for them: liability that flows down, an incident clock, and the audit criteria that already exist.

GOVERNS: WHAT THE PAPER SAYS. CONTROLS 11, 12.

WHAT THE NEWS SAID	WHAT THE RECORD SHOWS	THE ACTION
One vendor's stolen tokens reached 700 companies through integrations nobody had reviewed.	Third parties sit in 48 percent of breaches, and a public buyer now writes agent controls into its contract terms.	Put the first three actions in the contract and in the SOC 2 scope.

The fourth action puts the first three into contracts and audits. Management has read the liability clause for its three largest AI vendors and flowed its rules down to every contractor using AI on its work; a named person must be told of unauthorized use or harmful output within a stated time; and the organization's SOC 2 scope names its agents. These are the escalation triggers a board can see: what incidents reach it, and how fast.

Evidence

> The Universal Service Administrative Company published RFP IT-26-139 in September 2026 with a Privacy and Security Addendum. Section 3.3 says the "Contractor shall not use, implement, build, or deploy AI tools, services, or code of any type without prior written approval," requires an activity log kept to the buyer's retention policy, flows every rule down to

subcontractors, and requires the contractor to report unauthorized use or harmful output within one hour. The procurement office did not read the post-mortems; it reached the same list because the controls are the ordinary ones.

Verizon's 2026 Data Breach Investigations Report puts third parties in 48 percent of breaches, up 60 percent year over year. FINRA's Salesloft alert told firms to disconnect the integration, rotate credentials, audit logs, apply least privilege to third-party apps, and report to FINRA, the SEC, and the FBI.

USAC RFP IT-26-139, Addendum 3.3; Verizon 2026 DBIR, 19 May 2026; FINRA alert, Aug 2025

> "Contractor shall not use, implement, build, or deploy AI tools, services, or code of any type without prior written approval."
>
> UNIVERSAL SERVICE ADMINISTRATIVE COMPANY, RFP IT-26-139, PRIVACY AND SECURITY ADDENDUM 3.3, SEPTEMBER 2026

Implication

Commercial leaders feel this requirement before security leaders do. A buyer who writes Addendum 3.3 into terms is asking the fifteen questions in this brief, and an organization that can answer them is eligible for the work while one that cannot is disqualified before the evaluation begins. Governance here is a business enabler rather than a compliance exercise.

DECISION FRAMEWORK

Three boardroom test questions, and three steps to take

> "Which AI systems can act autonomously, and who is accountable when they get it wrong?"
>
> KHWAJA SHAIK, IBM CTO AND NACD DIRECTOR, LINKEDIN, SEPTEMBER 2026

Directors do not need to read the incident reports themselves, but they do need to be able to answer three questions about their own organization. Boardroom test questions include the following:

1. Who owns agent access, by name?

If the answer is a committee, a vendor, or the team that deployed each agent, the organization sits with the 93 percent of surveyed leaders who cannot name an accountable person. The evidence points to the chief information security officer, because identity, access, logging, and revocation already run on that function's systems. Business leaders decide which work agents do; the security function decides what each agent can reach.

2. Which committee receives the report, and how often?

The audit or risk committee already owns the SEC four-business-day disclosure clock, the insider-threat program, and third-party risk. A separate technology committee adds a body without adding a control. A quarterly cadence matches the revocation test and the annual breach survey cycle.

3. What three numbers does management bring?

Agents with write access to a system of record, with the executive signature that permits each. Incidents in the quarter in which an agent acted outside its approval, with the time from event to the named person being told. The revocation test: the date it was last run at the identity provider and the time to shutdown in minutes.

Three steps follow from the answers:

Assign one owner and ask for the inventory. Boards should name the owner of agent access this quarter and set a date for the first inventory. Questions 1, 7, and 15 of the diagnostic cannot be answered without the security function in the room, so taking it is the first meeting.

Route the report to the committee that already owns disclosure. Boards should place the three numbers on the audit or risk committee's quarterly agenda and codify the escalation trigger: which agent incidents reach the committee, and within what time.

Tie the four actions to the audit scope and the contracts. Management needs to bring agents into the SOC 2 scope at the next examination and flow the write rule, the data rule, and the incident clock down to every vendor and contractor using AI on the organization's work.

DECISION SUPPORT

Fifteen fact questions return the permission map.

Each question is a fact a chief information security officer, chief information officer, or chief AI officer can answer in under a minute. Yes is always the safe answer, and Partial counts as not in place. The result is which of eight common agent jobs an organization's controls clear today, which are one control away, and the one control to add first (Exhibit 7). The diagnostic belongs to Band 4 of the Intelligence Method, Adaptive Governance, where governance is the function that lets an organization "go farther and faster."

EXHIBIT 7

The fifteen controls, the action each belongs to, and the SOC 2 criterion an auditor tests for it

#	THE FACT QUESTION	ACTION	SOC 2 CRITERION
1	Can you produce today a list of every AI agent, tool, or automation holding credentials to a company system, with what each can read and write?	I	CC6.1 inventory
2	Does every AI tool or agent get a written approval from a named person before it runs on company systems or data?	I	CC6.2 authorization
3	Is there a written rule for which company data may enter which AI tool, and who authorizes exceptions?	I	CC6.1; CC6.7
4	Is there a written rule, signed by an executive, naming which decisions an agent may make on its own, which it may only support, and which systems of record it may write to?	II	CC6.3 role-based access
5	Is every agent that reads outside content either kept from writing or sending on internal systems, or required to get a human approval per action?	II	CC6.6 boundary
6	Is outbound internet access denied to agents by default, with an allowlist?	II	CC6.6 boundary
7	Can you revoke an agent's access from outside the agent, at the identity provider, and was that tested and timed in the last quarter?	II	CC6.3 removal of access
8	Are agent actions logged by the platform rather than by the agent?	III	CC7.2 monitoring
9	For every agent with a target metric, is there a counter-metric?	III	CC4.1 evaluations
10	Is every agent's output reviewed by a person or by a pass under a different instruction?	III	CC4.1 evaluations
11	If an agent does something unauthorized or produces harmful output, is there a named person who must be told, and within how long?	IV	CC7.3 incident response
12	For your three largest AI vendors and any contractor using AI on your work, do you know who is liable when the agent acts wrongly, and does your paper flow your rules down?	IV	CC9.2 vendor risk
13	For each AI tool in use, do you know the model version, its published limitations, the provider's training-data policy, and the knowledge cutoff, and are you told before the version changes?	III	CC8.1 change management
14	Is there a written rule for when a pilot moves from full human review of every output to sampling?	III	CC8.1 testing before change

#	THE FACT QUESTION	ACTION	SOC 2 CRITERION
15	Does the security team run agents of its own?	I	CC7.2 detection

Source: RBD. AI Agent Security Diagnostic, questions as published; AICPA Trust Services Criteria (2017, points of focus revised 2022), criterion numbers as commonly referenced. Derived from Band 4, Adaptive Governance, in The Intelligence Organization (Starkey, 2026).

Take the diagnostic

The fifteen questions take under fifteen minutes and return the permission map on screen: the agent jobs the organization's controls clear today, the jobs one control away and which control clears the most, the jobs that should wait and in what order, and the fifteen answers as a table to bring to the security function.

Take the AI Agent Security Diagnostic

The window for the first action

The rules, contract terms, and audit expectations around AI agents are changing on dated schedules: a public buyer's clauses this month, a state disclosure rule on 1 October 2026, the EU's high-risk obligations on 2 December 2027. Boards cannot assume that the oversight built for software vendors will cover systems that hold credentials and act on their own. Boards that assign the owner, receive the three numbers, and bring agents into the audit scope will deploy more of the read-heavy work agents already do well, and will do so with the accountability their insider-risk programs already provide. The time to assign the owner is this quarter.

Put the permission map in front of the board.

A focused working session takes the diagnostic result, the agent inventory, and the existing SOC 2 scope and produces the quarterly report an audit or risk committee can adopt at its next meeting.

SOURCES

ABOUT THE RESEARCH

This brief draws on 22 sources across five categories, read at the primary source between 18 and 20 September 2026: the forecast's own self-grading posts and an independent tracker of its 66 dated predictions; six incident investigations and vendor disclosures; three security surveys, each labeled where vendor-published; the regulators, courts, buyers, and standards bodies whose text sets the obligations; and one director's public statements. Survey correlations are reported as the share of breached organizations lacking a control at the time, never as a cause. The fifteen controls are the questions of the AI Agent Security Diagnostic, published by RBD. and derived from Band 4, Adaptive Governance, in *The Intelligence Organization*.

References

FORECAST RESEARCH AND SELF-GRADING

- AI Futures Project. "Grading AI 2027's 2025 Predictions." 12 February 2026. "Q1 2026 Timelines Update," 2 April 2026. "Q2.5 2026 Timelines Update: Uplift and Revenue," 16 August 2026.
- AI 2027 Tracker. All predictions, 66 dated claims with status labels. Independent; last updated 14 September 2026.

INCIDENT INVESTIGATIONS AND PRIMARY DISCLOSURES

- METR and Redwood Research. "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident." 26 August 2026. OpenAI and Hugging Face joint statement, 21 July 2026.
- FINRA. Cybersecurity alert on the Salesloft Drift AI supply chain incident, August 2025. Google Threat Intelligence attribution to UNC6395.
- AI Incident Database, incident 1152 (Replit), 18 July 2025.
- Anthropic. Agentic misalignment study, June 2025. Anthropic, November 2025, on a state-backed intrusion campaign run with a commercial coding agent.
- OX Security. "The Mother of All AI Supply Chains." 15 April 2026. The Register, 16 April 2026.
- Cloud Security Alliance. "MCP Security Crisis: Systemic Design Flaws in AI Agent Infrastructure." 4 May 2026.
- Snyk. ToxicSkills research, 5 February 2026. Vendor-published.

SECURITY AND BREACH SURVEYS

- IBM and Ponemon Institute. Cost of a Data Breach Report, 2025 edition (30 July 2025) and 2026 edition (29 July 2026). Vendor-published.
- Gravitee. State of AI Agent Security Report 2026, updated April 2026. 750 senior technology leaders, UK and US. Vendor-published.
- Verizon. 2026 Data Breach Investigations Report, news release 19 May 2026.

REGULATORS, BUYERS, AND STANDARDS BODIES

- Universal Service Administrative Company. RFP IT-26-139, Privacy and Security Addendum, section 3.3. September 2026.
- European Union. Regulation (EU) 2026/1744, the Digital Omnibus on AI, in force 27 July 2026. State of Connecticut, SB 5 (2026).
- US Securities and Exchange Commission, Release 33-11216 (2023). Delaware Supreme Court, Marchand v. Barnhill (2019).

- Model Context Protocol. "The MCP Registry," read 20 September 2026. Linux Foundation, Agentic AI Foundation announcement, 9 December 2025.
- AICPA. Trust Services Criteria (2017), revised points of focus (2022).

PRACTITIONER STATEMENTS AND RBD. RESEARCH

- Shaik, K. LinkedIn posts, September 2026. IBM CTO; NACD director. Practitioner view.
- Starkey, M. C. The Intelligence Organization. 2026. Band 4, Adaptive Governance.
- RBD. AI Agent Security Diagnostic, 2026. The full evidence file for this brief is the research brief How boards should govern AI agents in 2026.

Vendor-published sources are named as such. IBM's 92% is a share of the 21% of breached organizations whose own AI was breached, not of all breaches. Survey figures describe organizations lacking a control at the time of a breach, never a cause. SOC 2 criterion numbers follow the 2017 Trust Services Criteria as commonly referenced and should be confirmed against the AICPA text before any compliance decision.

contact@rbdco.ai