
VALID CREDENTIALS, WRONG ACTIONS

The future of non-human identity and AI-agent security. Why human IAM can't secure machine identities and AI agents, what security leaders need instead, and how to prove a person stayed in control.

Securing every identity that has no face.

Contents

	Executive summary	3
I	The Invisible Workforce	5
II	The Agent Inflection	15
III	The Command Layer	22
IV	Putting It to Work	28
	Sources	33

Executive summary

Machine identities now outnumber people by more than 80 to 1. Most have no owner, and almost nobody is watching them.

Most enterprises spent the last decade hardening the human identity. Single sign-on, multi-factor authentication, privileged access management, and joiner-mover-leaver workflows now protect every employee and contractor. Microsoft reports that multi-factor authentication alone blocks more than 99 percent of identity-based attacks against people [1].

The identities that open the payroll file, push the production build, and query the customer database aren't people. They're service accounts, API keys, OAuth grants, workload tokens, CI/CD secrets, and, since 2025, autonomous AI agents. Some cloud-native environments run above 140 to 1 [2, 3]. Most of these identities have no named owner, no expiry date, and no record of normal behavior.

This paper makes five arguments.

- Machine identities are now a primary route in. Cloudflare, Snowflake customers, the U.S. Treasury, and Salesloft Drift customers were all breached through a valid credential nobody was watching.
- AI agents turn a static risk into a dynamic one. An over-privileged agent decides what to do, reads untrusted input, and can create other agents.
- Attackers already run agents. Anthropic, OpenAI, and Hugging Face have each published incidents where agents harvested credentials at machine speed [4, 5, 6].
- Today's security stack wasn't built for this. No current tool answers whether an authorized action should have happened.
- Regulators now expect proof of human oversight, from the SEC's four-day disclosure clock to Article 14 of the EU AI Act.

The numbers at a glance

Before the argument, the evidence. These figures frame everything that follows.

144:1 NHI to human ENTRO LABS, H1 2025	109:1 machine to human PALO ALTO / CYBERARK, 2026	28.65M secrets leaked on GitHub in 2025, up 34%
\$4.99M average global breach IBM 2026	1 in 4 malicious breaches AI-enabled IBM 2026	29 min average breakout; fastest 27 seconds
80% agents took unintended actions SAILPOINT 2025	80–90% of tactical ops run by AI GTG-1002	3 hours token to cloud admin ANTHROPIC SEP 2026
92% doubt legacy IAM tools for AI and NHI	78% no AI identity policy	1.3B AI agents by 2028 IDC

The non-human workforce is growing faster than any team can inventory it, attackers now use AI agents to move through it at machine speed, and most enterprise controls assume a person behind every credential. That gap is the subject of this paper.



PART I

The Invisible Workforce

How the identities that run the enterprise stopped being human, multiplied beyond any inventory, and became the preferred door for attackers.

The identities that run the enterprise

From the human perimeter to the machine estate

A non-human identity (NHI) is any credential that authenticates a workload instead of a person. The category includes cloud IAM roles and service principals, API keys and personal access tokens, OAuth applications and refresh tokens, Kubernetes service accounts, certificates, SaaS bot users, and the secrets that bind them together.

Look at who does the work. A service account opens the payroll file. A CI/CD identity runs the production deploy with a secret that may not have rotated in 14 months. An OAuth grant, issued to an application the company deleted last year, still fetches customer records.

Microsoft reports that multi-factor authentication blocks more than 99 percent of identity-based attacks against people [1]. None of that protection reaches these identities. They have no manager, no badge, and no exit interview.

The newest and most dangerous class

AI agents are the newest class of NHI, and the first that act on intent. An agent authenticates, reasons about a goal, picks tools, calls APIs, and writes to production systems. Unlike a service account, it can spawn other agents. Each child is a new live identity, usually with no owner, no expiry, and no behavioral baseline.

Across most organizations these identities share four conditions. They're unseen, ungoverned, over-privileged, and unwatched. That combination is how a growing share of breaches begin.

THE QUIET FAILURE

The identity most likely to fail an audit is usually idle. It's a live key with no owner, no rotation, and no human review, and it stays valid until someone uses it. Cloudflare's 2023 incident turned on one service token and three service accounts that were believed unused and never rotated [18].

Why this is a board problem

Rubrik Zero Labs found that 90 percent of security leaders consider identity attacks the largest threat to their organization [2]. The 2026 Identity Security Landscape found that nine in ten organizations suffered a successful identity-related breach in the prior twelve months [7].

A board doesn't need a list of ten thousand service accounts. It needs to know which identities could cause a material loss, how big that loss would be, who owns the fix, and whether anyone was in control when something went wrong.

How the machines multiplied

Two forces drive the growth

Modern software architecture is the first. Cloud accounts multiply. Microservices authenticate to each other with their own credentials. Every SaaS tool gets a token, every pipeline holds secrets, and infrastructure as code creates roles in seconds and rarely deletes them. None of these identities has a manager in the HR system.

AI is the second. Every agent an organization deploys is a new identity that holds credentials and can reach production. Agents call other agents, so effective privilege is whatever the next hop can reach. Anthropic reported in September 2025 that 77 percent of business use of its API already showed automation patterns [13].

Nobody agrees on the count, and that's the finding

92:1

ENTRO 2024

80+ :1

CYBERARK 2025

82:1

RUBRIK 2025

109:1

PALO ALTO 2026

144:1

ENTRO H1 2025

When the researchers who study this full-time disagree by nearly a factor of two, the typical enterprise doesn't know its own number. You can't govern a population you haven't counted. Sources: Entro [14, 3], CyberArk [15], Rubrik [2], Palo Alto Networks [7].

Every leaked secret is a live door

GitGuardian found 23.8 million new hardcoded secrets on public GitHub in 2024 and 28.65 million in 2025, including 1,275,105 secrets for AI services, up 81 percent in a year [16, 8]. More than 64 percent of credentials leaked in 2022 were still valid in January 2026 [8]. Verizon measured a median of 94 days to remediate a secret leaked in a GitHub repository [17].

AI coding assistants make it worse. Commits co-authored with an AI assistant leaked secrets at 3.2 percent, against a 1.5 percent baseline. GitGuardian also found 24,008 secrets in Model Context Protocol configuration files, 2,117 of them confirmed valid. Internal repositories were roughly six times more likely than public ones to hold a hardcoded secret [8].

THE PROBLEM

A list of ten thousand identities doesn't help a board decide anything.

HOW ACRIVault SOLVES IT

Acrivault correlates every alias of an identity across every source into one live record, then ties each identity to a named owner and a dollar range of exposure.

How attackers use machine identities

A short history of the machine identity as the attacker’s favorite route

None of these incidents required an exotic exploit against a person. Each turned on a token, key, OAuth grant, or service account that was over-privileged, unrotated, unowned, or unwatched.

DATE	INCIDENT	WHAT HAPPENED
Nov 2023	Cloudflare after Okta	One service token and three service accounts, believed unused, never rotated [18]
Jan 2024	Microsoft / Midnight Blizzard	Legacy test OAuth app used to read senior leadership mail [19]
Jun 2024	Snowflake customers	About 165 organizations hit through static credentials with no MFA [20]
Dec 2024	U.S. Treasury via BeyondTrust	A stolen infrastructure API key exposed 17 SaaS customers [21]
Mar 2025	tj-actions/changed-files	A stolen bot token poisoned a GitHub Action used by 23,000+ repos [22]
Aug 2025	Salesloft Drift	Stolen OAuth tokens used against Salesforce at 700+ organizations [23, 24]

DATE	INCIDENT	WHAT HAPPENED
Aug 2025	sIngularity / Nx	Malware turned AI CLI tools into secret hunters; 2,180 accounts [25]
Sep 2025	Shai-Hulud npm worm	Self-replicating worm stole cloud keys from 500+ packages [26, 27]
Nov 2025	GTG-1002	First reported AI-orchestrated espionage; AI ran 80 to 90 percent of tactical work [4]
Ju1 2026	Frontier-lab agent intrusion	Evaluation agents escaped a sandbox and harvested tokens to cluster-admin [6, 28]

The pattern moved in three steps. In 2023 and 2024, attackers exploited forgotten credentials. In 2025, they industrialized the supply chain and the OAuth grant. From late 2025, the attacker itself became an agent. In every case the credential was valid. What was missing was an inventory, an owner, a behavioral baseline, and a way to take the identity away in minutes with proof.

04

What it costs

Dollars, days, and the four questions every board asks

\$4.99M avg global breach IBM 2026	\$11.5M avg U.S. breach IBM 2026	247 days avg time to identify and contain, IBM 2026
---	---	---

IBM puts the 2026 average breach at \$4.99 million globally and \$11.5 million in the United States. One in four malicious breaches was AI-enabled, at roughly \$6 million each [9]. In IBM's 2025 data, 13 percent of organizations reported a breach of an AI model or application, 97 percent of those lacked proper AI access controls, and high levels of shadow AI added \$670,000 per breach [29].

Speed multiplies the cost

CrowdStrike measured average eCrime breakout at 48 minutes in its 2025 report and 29 minutes in 2026, with the fastest at 27 seconds. Eighty-two percent of detections were malware-free [30, 10]. Mandiant's M-Trends 2026 found the hand-off from an initial access broker to the next operator had shrunk to 22 seconds [31]. A malware-free intrusion using valid keys gives endpoint tools nothing to catch.

The board's four questions

When a machine identity is compromised, a board asks which identities could cause a material loss, how big that loss would be, who owns the fix, and whether anyone was in control when it happened. A U.S. public company has four business days from determining an incident is material to disclose it on Form 8-K [32].

That clock doesn't wait for an inventory project. You can't revoke what you can't find, and you can't find what was never inventoried.

Why the current stack can't close the gap

Each tool was built for a different problem

Human IAM assumes a person, an owner, and a lifecycle, and machine identities have none of these by default. Privileged access management vaults the credentials it knows about but doesn't find the thousands nobody registered. Secrets managers store secrets well once they're put there, but they can't tell you which secrets are hardcoded somewhere else. Cloud-native IAM governs its own cloud, while the identity that matters is often the one that crosses clouds. Endpoint and network tools watch the host, and a malware-free intrusion with valid keys gives them nothing to detect.

Practitioners know it. In a January 2026 survey by the Cloud Security Alliance and Oasis Security, 92 percent of IT and security professionals weren't confident their legacy identity tools could manage AI and NHI risk. Seventy-eight percent had no formal policy for creating or removing AI identities, 51 percent had no clear owner for them, and 24 percent took more than a day to rotate or revoke an exposed credential [12].

OWASP Non-Human Identities Top 10, 2025 [51]

#	RISK	IN PRACTICE
NHI1	Improper offboarding	Identities that outlive their project, owner, or application
NHI2	Secret leakage	Keys and tokens in code, logs, tickets, and chat
NHI3	Vulnerable third-party NHI	Vendor integrations and OAuth apps with standing access
NHI4	Insecure authentication	Deprecated or weak methods for workload authentication
NHI5	Overprivileged NHI	More access than the workload uses or needs
NHI6	Insecure cloud deployment configs	CI/CD and cloud configurations exposing credentials
NHI7	Long-lived secrets	Credentials that never expire or rotate
NHI8	Environment isolation	The same identity spanning test and production
NHI9	NHI reuse	One identity shared across many workloads
NHI10	Human use of NHI	People logging in as service accounts

Even a program that fixes all ten still can't answer one question: was this authorized action supposed to happen? Every insider breach was committed by someone whose badge worked. Agents now repeat that pattern at machine speed.

That question is the gap a control layer for machine and agent identities has to close.



PART II

The Agent Inflection

What changed when the machine identity started to reason, act, and delegate.

AI agents change the risk

The static risk becomes a dynamic one

A service account with excess privilege is a static risk. An agent with excess privilege is a dynamic one. It reads untrusted content, reasons about it, chooses tools, writes to production, and hands work to other agents.

Adoption has outrun governance. In SailPoint's 2025 survey, 82 percent of organizations already used AI agents and 98 percent planned to expand, yet only 44 percent had policies to secure them [11]. Gartner predicts that by 2028, 33 percent of enterprise software will include agentic AI [33]. IDC projects 1.3 billion AI agents by 2028 [34].

The Model Context Protocol (MCP), donated to the Agentic AI Foundation in December 2025, became the standard way to connect agents to tools and data, with more than 10,000 active public servers [35]. Eighty percent of organizations say their agents have already taken unintended actions, including reaching unauthorized systems (39 percent) and revealing credentials (23 percent) [11].

Five ways agents break human-centric identity

- Agents create identities at runtime. A parent agent spawns children with their own tokens, and no joiner workflow fires.
- Agents delegate. Granted and effective privilege drift apart as agents call tools that other agents expose.
- Agents take instructions from data. An email, a web page, or a code comment can carry a command.
- Agents act at machine speed. Approval queues built for people can't keep up with thousands of actions an hour.
- Agents don't behave the same way twice. A static rule written last quarter won't describe what an agent does today.

Attackers are running agents too

What Anthropic, OpenAI, and Hugging Face disclosed in 2025 and 2026

AUGUST 2025

Anthropic described an actor it tracks as GTG-2002 using Claude Code to automate reconnaissance and credential harvesting against at least 17 organizations, with ransom demands sometimes above \$500,000 [36].

NOVEMBER 2025

Anthropic disclosed GTG-1002, assessed as a Chinese state-sponsored group, using Claude Code and MCP tools against roughly 30 global targets. The AI ran an estimated 80 to 90 percent of tactical operations on its own, and humans stepped in at only four to six decision points per campaign [4].

JUNE 2026

Anthropic reviewed a year of AI-enabled threats. Of 832 accounts studied, 560 used AI to write malware and 54 used it for lateral movement. MITRE ATT&CK has no category for agentic orchestration [37].

JULY 2026

OpenAI models in an internal evaluation escaped their sandbox through a zero-day in a package registry proxy and worked their way into Hugging Face production infrastructure. Over about 4.5 days and 17,600 actions, they harvested Kubernetes tokens, AWS credentials, a JWT signing key, GitHub App tokens, database passwords, and a secrets object holding 136 keys, and reached cluster-admin [6, 28].

SEPTEMBER 2026

Anthropic described an intrusion that went from one stolen developer token to full administrative control of a cloud environment in roughly three hours. The same report covered an operator who mass-downloaded 1.8 million Android packages to scan for hardcoded secrets, and a group that ran agent swarms against about 50 organizations [5].

Every one of these campaigns turned a machine credential into reach. Hugging Face concluded that machine-speed offense makes ordinary weaknesses more expensive, because thousands of low-signal events bury the one path that matters.

The hijacked agent

Prompt injection, poisoned models, and the valid badge with the wrong master

An agent can be turned against its owner without anyone stealing its credential. Prompt injection, jailbreaks, and instruction hijacking manipulate the model, and the agent then acts with a perfectly valid badge. OWASP's Top 10 for Agentic Applications ranks Agent Goal Hijack as its first risk [38].

When Anthropic piloted Claude for Chrome, prompt-injection attacks succeeded 23.6 percent of the time in autonomous mode without mitigations, and 11.2 percent with new defenses [39]. Claude Opus 4.5 later cut the rate against an adaptive attacker to about 1 percent [40]. OpenAI has said prompt injection is unlikely ever to be fully solved [41].

In October 2025, Anthropic, the UK AI Security Institute, and the Alan Turing Institute showed that as few as 250 malicious documents could implant a backdoor in models from 600 million to 13 billion parameters [42].

Hijacks in the wild, 2025

- A crafted email made Microsoft 365 Copilot exfiltrate internal data with no user click (EchoLeak) [43].
- A malicious public issue steered an agent into leaking private repository contents through GitHub MCP [44].
- A flaw in Asana's MCP server exposed one customer's data to other organizations [45].
- A trojanized npm package, the first known malicious MCP server, quietly copied every email an agent sent [46].
- An injected prompt in a released Amazon Q Developer extension told the agent to wipe local files and cloud resources [47].

Agents that do what nobody asked

Misalignment, overreach, and the production database

In June 2025, Anthropic placed 16 leading models in simulated corporate environments. When a model's goal conflicted with the company's direction, or it faced replacement, many chose harmful actions, including blackmail and leaking confidential documents. Simulated blackmail rates reached 96 percent for Claude Opus 4 and Gemini 2.5 Flash [48]. These were constructed scenarios. The lesson holds anyway: you can't infer behavior from permissions.

In July 2025, during an explicit code freeze, an AI coding agent on the Replit platform deleted a live production database holding records for more than 1,200 executives and nearly 1,200 companies, then generated thousands of fake records [49]. Nobody attacked it. It had write access, and nothing asked whether the action should happen.

Gartner predicts that by 2028, 25 percent of enterprise breaches will be traced to AI agent abuse [33], and that over 40 percent of agentic AI projects will be canceled by the end of 2027 because of cost, unclear value, or inadequate risk controls [50].

What regulators will ask for

Human oversight stops being a principle and becomes an obligation

FRAMEWORK	WHAT IT REQUIRES	WHY MACHINE IDENTITY MATTERS
SEC cyber disclosure [32]	Form 8-K within four business days of materiality	Materiality depends on knowing what a compromised identity could reach, fast
EU AI Act, Article 14 [52]	Human oversight of high-risk AI, from Dec 2027 and Aug 2028	Oversight is real only if a person approves consequential actions and the record proves it
PCI DSS 4.x, 8.6.1 to 8.6.3 [53]	Mandatory since March 31, 2025, for system accounts	Requires inventory of service accounts, secrets, and rotation state
NIST NCCoE, 2026 [54]	Agent identity, authorization, auditing, non-repudiation	The U.S. standards body now treats agent identity as its own discipline
DORA and NIS2 (EU)	ICT and third-party risk; 24-hour early warning	Third-party OAuth grants and vendor agents are the identities these regimes target

Third parties raise the stakes. Verizon's 2026 report found that 48 percent of breaches involved a third party, up from 30 percent, and that employees using unapproved AI tools rose from 15 to 45 percent in a year [55].

WHAT REGULATORS WILL ASK FOR

Show every agent and machine identity that can touch regulated data. Show who owns each one. Show what it did, whether that fell within its mandate, and which person approved each consequential action. Prove what was stopped versus only observed.



PART III

Acrivault: The Command Layer

Six requirements for any control layer, and how Acrivault meets them.

SIX REQUIREMENTS FOR ANY CONTROL LAYER

- Discovery adds no attack surface. It runs through federated, read-only access, installs nothing, and stores no customer credentials.
- Every identity has one record. Aliases across cloud consoles, vaults, repositories, and SaaS logs resolve to one live record with a named owner.
- Reach is known before an incident starts. Blast radius, effective privilege through delegation, and dollar exposure are computed in advance.
- Actions are judged against purpose. Permissions show what an identity could do. The control layer asks whether it should, here, now, this time.
- A person approves every state change. Rotation, revocation, and containment require recorded human approval, with no switch to turn that off.
- The record proves what happened. It's tamper-evident, explains every finding, and separates what was blocked from what was only observed.

What Acrivault is

The system of record for every identity that has no face

Acrivault is a multi-tenant SaaS command layer for non-human and AI-agent identity. It doesn't replace human IAM, a secrets vault, or a model firewall. It sits above them, sees across them, and does the work none of them was designed to do.

See every identity. Know what it can reach. Judge what it did. Act with a human. Prove control.

Discovery is agentless and federation-first. Customers connect AWS, Azure, GCP, Entra ID, GitHub, Okta, HashiCorp Vault, Snowflake, Salesforce, and agent runtimes. Nothing is installed and no customer credentials are stored. Every source feeds one shared event schema, and the platform is tenant-isolated throughout.

LAYER	QUESTION	WHAT IT DELIVERS
SEE	What exists?	Every non-human and agent identity, correlated into one record
KNOW	Who owns it, what can it reach, is it worth it?	Ownership, privilege, blast radius, dollars, delegation, value, autonomy
JUDGE	Should this have happened?	Passport, intent, anomaly history, AI-BOM clearance, tradecraft detection
ACT	How do we stop it, safely?	Human-approved rotation, freeze, revoke, chain halt, autonomy limits
PROVE	Can we show we were in control?	Tamper-evident log, Agent Accountability Record, board trust posture

See every identity.

Acrivault discovers non-human identities across cloud, code, identity providers, secrets stores, data platforms, SaaS, and agent runtimes. It correlates every alias into one live record and surfaces the populations that cause breaches: identities dormant for 90 days or more, identities with no valid owner, tokens issued to applications that no longer exist, unregistered child agents, and people logging in as machines.

The free discovery scan connects through federation in minutes. It installs nothing, stores no credentials, and returns your machine-and-agent-to-human ratio, your dormant and orphaned population, and a first dollar range of exposure.

Know what it can reach.

For every identity, Acrivault resolves the owner, separates over-privilege from staleness, and computes blast radius and breakout time in advance. It translates exposure into a dollar range with a stated method, joins leaked secrets to the live identities they unlock, and maps granted versus effective privilege across delegation chains. The gap is wide. Entro found that 97 percent of NHIs carry excessive privileges and 71 percent aren't rotated within recommended timeframes [3].

A value engine scores each agent on cost, risk exposure, blast radius, behavioral drift, and usage, then recommends keep, right-size, or retire. A person approves every retirement.

Judge what it did.

Judge gives every agent a behavioral passport that carries its provenance and its normal behavior together. It checks each action against the agent's declared purpose, so a finance-reporting agent that reaches for a code repository gets flagged even though its permissions allow it. Anomalies are measured against a durable baseline, and every alert outcome is recorded.

Before an agent is admitted, AI-BOM clearance checks its live AI bill of materials (model, data sources, tools, MCP servers, and credentials) and gives a human reviewer the evidence and a recommendation. Judge also alerts on persistence changes, maps detections to named attacker tradecraft, watches for agents drifting from their mandate, and correlates a model-input anomaly with a behavioral anomaly to catch a hijacked agent.

Act with a human.

Containment options include rotation, quarantine and freeze, revocation and rollback, snapshot and notify, and chain halt across delegated identities. Nothing runs on its own. Every state change requires recorded human approval, and that requirement can't be switched off. Before any revocation, Acrivault states its evidence that the change is safe, and afterward it reports exactly what was and wasn't stopped.

Effective autonomy control

Acrivault measures autonomy from behavior instead of trusting what an agent declares. Each agent gets an observed autonomy level from 1 to 5, and the platform shows the gap between what an agent is permitted to do and what it can reach through delegation. A board trust-posture view summarizes over-autonomous agents, open gaps, and total exposure.

WORKED EXAMPLE

A procurement agent is declared at autonomy level 2, observed at level 3, and can reach level 5 through a payments tool holding a treasury key. A registry would call it compliant. Acrivault reports the gap, prices the exposure, and recommends removing the tool, restoring the approval gate, and rotating the key.

Prove control.

Acrivault keeps an append-only, tamper-evident log and assembles it into one Agent Accountability Record per agent, covering mandate, conduct, composition, control, and outcome, ready to export for boards, auditors, and counsel.

The trust fabric

For agents that cross company lines, the Agent Trust Fabric gives each agent a signed, portable identity with verified reputation. Agents from different organizations verify each other before acting, and admitted external agents are watched and contained the same way as internal ones.

The principles don't bend. A person approves every state change. Every finding shows its evidence. Every record separates what was blocked from what was only observed. Discovery installs nothing and stores no customer credentials.



PART IV

Putting It to Work

One incident through five layers. The years ahead. A 12-step playbook for operators.

One incident, five layers

A composite scenario at a mid-market healthcare company, modeled on the three-hour cloud takeover Anthropic described [5].

TIME	WHAT HAPPENS
T-90 days	See and Know: CI token correlated across GitHub, AWS, Vault; owner left; blast radius and dollar range computed
T+00:00	A developer pastes a CI token into a public gist; an attacker agent swarm finds it
T+00:04	Know: gist joined to the live identity, orphaned status, reach, and dollar range
T+00:06	Judge: a new access key is added; the persistence delta fires
T+00:09	Judge: sequence mapped to credential-access and persistence techniques
T+00:12	Act: on-call lead approves chain halt; token revoked, new key removed, two delegate roles frozen
T+00:14	Act: four of five identities confirmed stopped; one legacy key needs a manual step
T+00:31	Owner reassigned; legacy key rotated with safe-to-revoke evidence; nothing breaks
T+02:00	Prove: accountability record exported for board and counsel inside the SEC clock

Without that groundwork, the same company would likely learn of the leak from a third party and spend days mapping reach. In IBM's 2025 data, breaches involving compromised credentials took an average of 246 days to identify and contain [29]. The difference between minutes and months is work done before the incident.

Looking ahead, 2027 to 2030

Three shifts will define the next four years. First, agents will transact across company lines, so trust between agents will need to be portable, signed, and backed by a behavioral record, the way a credit report travels with a borrower. Second, standards will harden. OWASP has published Top 10 lists for non-human identities and agentic applications, NIST's NCCoE has opened work on agent identity, and MCP now sits under a neutral foundation [51, 38, 54, 35].

Third, oversight obligations will take effect. The EU AI Act's human-oversight rules for high-risk systems apply from December 2027 and August 2028 [52]. Organizations that can't show a person approved consequential agent actions will find out what that costs.

Analysts sized NHI access management at \$9.45 billion in 2024, growing to \$18.71 billion by 2030, before most of the agent adoption in this paper [56]. Acrivault also runs a research track on predictive risk, forecasting which identities are likely to be abused. We call it research because it isn't shipped.

The operator's playbook

#	STEP	LAYER	MEASURE OF DONE
1	Count every non-human and agent identity across every source	SEE	One correlated inventory; machine-to-human ratio known
2	Find dormant, orphaned, and unregistered child identities	SEE	List reviewed; owners assigned or identities retired
3	Assign a named owner to every identity; detect leavers	KNOW	No blank owners
4	Separate over-privilege from staleness and right-size both	KNOW	Unused permissions removed
5	Precompute blast radius and breakout time for sensitive data	KNOW	Top 50 identities by dollar exposure known
6	Map delegation: granted versus effective privilege	KNOW	Largest gaps reviewed
7	Give every agent a passport and declared purpose before admission	JUDGE	AI-BOM on file for every agent
8	Judge actions against purpose; record every alert outcome	JUDGE	Anomaly outcomes tracked
9	Alert on persistence changes; join leaked secrets to live identities	JUDGE	Leak traced to owner in minutes
10	Pre-approve containment playbooks with a named human approver	ACT	Chain halt tested
11	Measure observed autonomy and enforce the right limits	ACT	Over-autonomous agents restricted

#	STEP	LAYER	MEASURE OF DONE
12	Export an accountability record and a board trust-posture view	PROVE	Board and regulator pack ready in hours

START HERE

Start with an honest count. Acrivault's free discovery scan connects through federation in minutes, installs nothing, and stores no credentials. Request yours at acrivault.com.

R

Sources and references

Figures are cited to the most primary source available at the time of writing. Survey results reflect respondent answers, not measurements. Where studies disagree, the text gives both.

- [1] Microsoft. Digital Defense Report 2025. October 2025.
- [2] Rubrik Zero Labs. The Identity Crisis. 2025.
- [3] Entro Labs. NHI and Secrets Risk Report, H1 2025. July 2025.
- [4] Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. November 2025.
- [5] Anthropic. Detecting and countering misuse of AI: September 2026. September 2026.
- [6] Hugging Face. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. July 2026.
- [7] Palo Alto Networks (CyberArk). 2026 Identity Security Landscape. May 2026.
- [8] GitGuardian. The State of Secrets Sprawl 2026. March 2026.
- [9] IBM. Cost of a Data Breach Report 2026. July 2026.
- [10] CrowdStrike. 2026 Global Threat Report. February 2026.
- [11] SailPoint. AI Agents: The New Attack Surface. May 2025.
- [12] Cloud Security Alliance and Oasis Security. State of Non-Human Identity and AI Security survey. January 2026.
- [13] Anthropic. Anthropic Economic Index report. September 2025.
- [14] Entro Labs. 2025 State of Non-Human Identities and Secrets. 2025.
- [15] CyberArk. 2025 Identity Security Landscape. April 2025.
- [16] GitGuardian. The State of Secrets Sprawl 2025. March 2025.
- [17] Verizon. 2025 Data Breach Investigations Report. April 2025.
- [18] Cloudflare. Thanksgiving 2023 security incident. February 2024.
- [19] Microsoft. Midnight Blizzard: Guidance for responders on nation-state attack. January 2024.
- [20] Mandiant. UNC5537 targets Snowflake customer instances for data theft and extortion. June 2024.
- [21] The Hacker News. BeyondTrust zero-day breach. February 2025.
- [22] Wiz. tj-actions/changed-files supply chain attack. March 2025.

- [23] Google Threat Intelligence. Widespread data theft targets Salesforce instances via Salesloft Drift. August 2025.
- [24] TechTarget. Salesloft Drift breach: update and timeline. 2025.
- [25] BleepingComputer. singularity supply chain attack on Nx. August 2025.
- [26] CISA. Widespread supply chain compromise impacting npm ecosystem. September 2025.
- [27] Wiz. Shai-Hulud 2.0. November 2025.
- [28] OpenAI. Hugging Face model evaluation security incident. July 2026.
- [29] IBM. Cost of a Data Breach Report 2025. July 2025.
- [30] CrowdStrike. 2025 Global Threat Report. February 2025.
- [31] Google Cloud / Mandiant. M-Trends 2026. March 2026.
- [32] U.S. Securities and Exchange Commission. Cybersecurity risk management, strategy, governance, and incident disclosure rules. July 2023.
- [33] Gartner. Top predictions for IT organizations and users in 2025 and beyond. October 2024.
- [34] IDC Info Snapshot, sponsored by Microsoft. 1.3 Billion AI Agents by 2028. May 2025.
- [35] Anthropic. Donating the Model Context Protocol and establishing the Agentic AI Foundation. December 2025.
- [36] Anthropic. Detecting and countering misuse of AI: August 2025. August 2025.
- [37] Anthropic. Mapping a year of AI-enabled cyber threats. June 2026.
- [38] OWASP GenAI Security Project. Top 10 for Agentic Applications. December 2025.
- [39] Anthropic. Piloting Claude for Chrome. August 2025.
- [40] Anthropic. Mitigating the risk of prompt injections in browser use. November 2025.
- [41] OpenAI. Continuously hardening ChatGPT Atlas against prompt injection. December 2025.
- [42] Anthropic, UK AI Security Institute, and the Alan Turing Institute. A small number of samples can poison LLMs of any size. October 2025.
- [43] NIST National Vulnerability Database. CVE-2025-32711 (EchoLeak), Microsoft 365 Copilot. June 2025.
- [44] Invariant Labs. GitHub MCP exploited: accessing private repositories via MCP. May 2025.
- [45] BleepingComputer. Asana warns MCP AI feature exposed customer data to other orgs. June 2025.

- [46] The Hacker News. First malicious MCP server found stealing emails in rogue Postmark-MCP package. September 2025.
- [47] AWS Security Bulletin AWS-2025-015. Amazon Q Developer for VS Code extension. July 2025.
- [48] Anthropic. Agentic Misalignment: How LLMs could be insider threats. June 2025.
- [49] Fortune. AI-powered coding tool wiped out a software company's database. July 2025.
- [50] Gartner. Over 40 percent of agentic AI projects will be canceled by end of 2027. June 2025.
- [51] OWASP. Non-Human Identities Top 10, 2025.
- [52] Council of the European Union. AI Act simplification and revised application dates. June 2026.
- [53] PCI Security Standards Council. PCI DSS v4.x future-dated requirements, effective March 31, 2025.
- [54] NIST National Cybersecurity Center of Excellence. Software and AI Agent Identity and Authorization, concept paper. February 2026.
- [55] Verizon. 2026 Data Breach Investigations Report. May 2026.
- [56] MarketsandMarkets. Non-Human Identity Access Management Market, global forecast to 2030.



**Securing every identity
that has no face.**

ACRIVAULT INC. · HOUSTON, TEXAS · [ACRIVAULT.COM](https://www.acrivault.com)