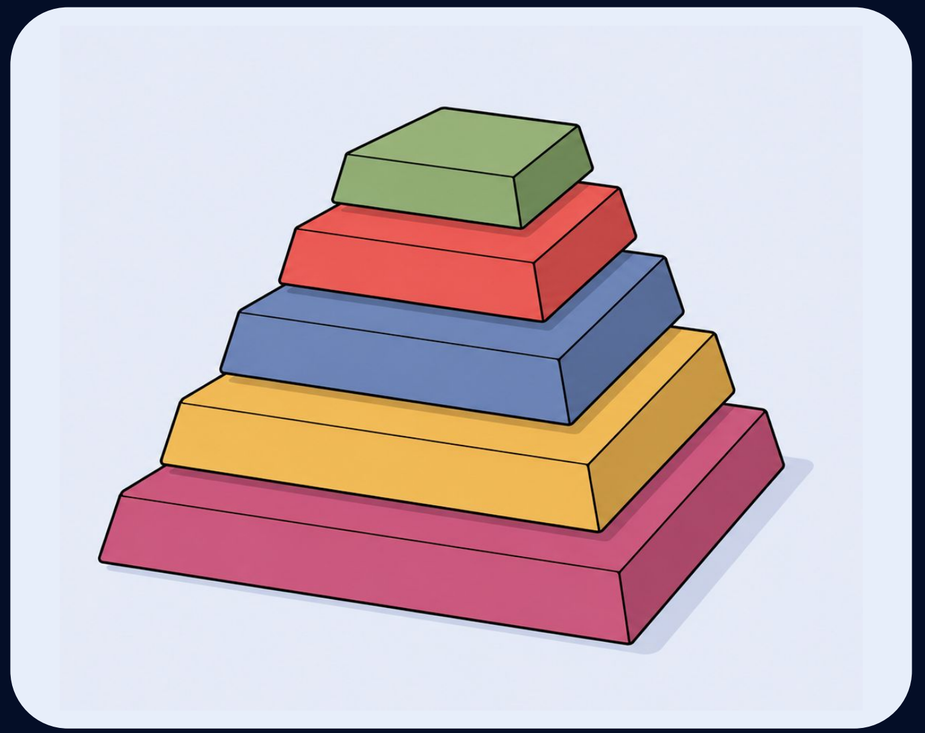
Everything comes down to infrastructure & identity security

AI security is no exception



Amber Bennoui · Sana Talwar

DAY OF SHECURITY



Amber Bennoui

Cybersecurity and AI product leader, 15+ years across F5 Networks, Datadog, Chainguard and DataRobot. Co-founder of the AI Security Alliance (AISECA). Researches AI supply chain security.



Sana Talwar

Product security engineer at ServiceNow, working on AppSec and AI-integration security. Previously a cybersecurity engineer at Johns Hopkins APL. WiCyS Los Angeles affiliate board.

JUNE 2026 · THE MOST POPULAR AI AGENT FRAMEWORK ON NPM

One dormant account. Nineteen minutes.

EARLY 2025

A contributor account, ehindero, goes quiet. Its publish rights on the scope stay.

16 JUN

Clean decoy published: easy-day-js 1.11.21

17 JUN · 01:01 UTC

easy-day-js 1.11.22 ships a postinstall dropper.

01:20 UTC

mastra 1.13.1 and 140+ packages republished with it, tagged latest.

19 JUN

Microsoft attributes it to Sapphire Sleet (North Korea).

The payload: a cross-platform stealer that went after 166 crypto wallet extensions and browser history.

Microsoft Threat Intelligence, 17 Jun 2026 (updated 19 Jun) · Snyk · StepSecurity · OSV MAL-2026-5960

That wasn't an AI exploit. It was offboarding.

WHAT THE HEADLINE SAID

Nation-state supply chain attack on the most popular AI agent framework.

WHAT ACTUALLY FAILED

A former contributor still had publish rights. The registry still took a manual publish.

Most of 2026's damaging AI incidents are **twenty-year-old identity failures** wearing an AI costume.

Everyone already agrees identity matters for agents.

FEB 2026

NIST CAISI

AI Agent Standards Initiative. Agent identity and authorization is a core pillar.

MAR 2026

CoSAI

Agentic IAM framework: agents as first-class identities, short-lived task-bound credentials.

> 50%

Gartner

of successful attacks on AI agents through 2029 will exploit access control issues.

ASI03

OWASP

Identity and Privilege Abuse is a top-three risk in the Agentic Top 10.

So why do the incidents keep happening?

NIST, 17 Feb 2026 · CoSAI Workstream 4, approved 20 Mar 2026 · Gartner, Wah & Senf, 20 Feb 2026 (via Astrix) · OWASP Top 10 for Agentic Applications

PART ONE

Five incidents, five controls you already own

01

Credential lifecycle

JUN 2026

Mastra npm

02

Authentication

APR–MAY 2026

Meta AI support bot

03

Authorization

JUN 2026

Copilot SearchLeak

04

Provenance

JUN 2026

A skill with no malware

05

Re-approval

FEB–JUN 2026

Agent config files

...plus one problem none of these controls covers.

INCIDENT 1 OF 5 · CREDENTIAL LIFECYCLE

A credential that outlived its owner

THE CONTROL

Joiner, mover, leaver. Publish rights are access, and access ends when the role does.

WHY IT FAILED

A dormant contributor kept publish rights on the whole scope.

Trusted publishing existed but wasn't required, so a manual publish with no provenance went through.

THE BORING FIX

Offboard publishers like employees.

Trusted publishing only, with every other path disabled.

Consumers: require provenance, ignore install scripts, wait a few days before adopting new versions.

The AI angle: agent frameworks land in environments that hold model API keys and tool credentials. Same failure, bigger prize.

Microsoft, 17 Jun 2026 · Snyk, 'A forgotten contributor account compromised the entire Mastra npm package scope'

INCIDENT 2 OF 5 · AUTHENTICATION

The support bot that handed over 20,225 accounts

1 Attacker connects from near the victim's location

→

2 Starts a password reset on Instagram

→

3 Asks Meta's AI support assistant to add a new email

→

4 Reset link goes to the attacker's email

→

5 Account taken over and defaced

Missing check: does this person own the account?



20,225

accounts, Meta's upper bound

17 APR → 31 MAY

attack window, 2026

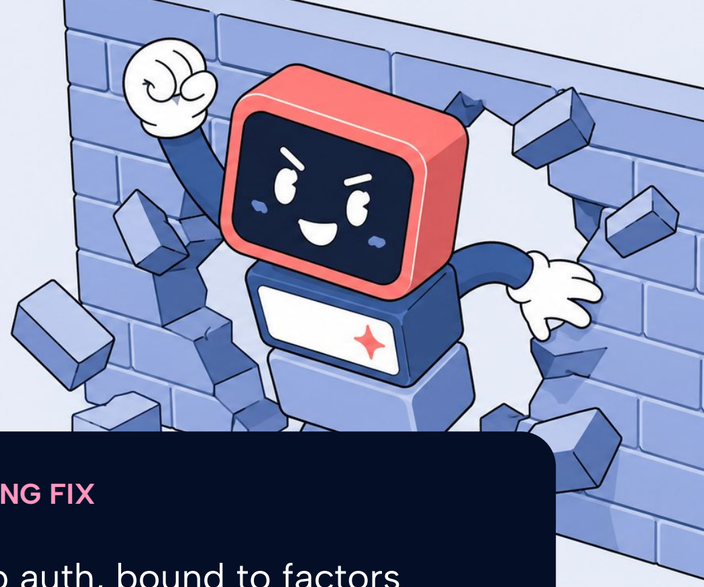
Obama White House · Sephora

among the accounts hit

Meta notice to Maine AG · Krebs on Security, Jun 2026 · Help Net Security, 8 Jun 2026

INCIDENT 2 OF 5 · AUTHENTICATION

A bot with write access and no identity check



THE CONTROL

Authenticate before an irreversible action. Changing a recovery channel is the most sensitive write an account has.

WHY IT FAILED

The assistant could change the recovery email and never verified who it was acting for.

Meta's own filing: the system didn't verify the email matched the account.

THE BORING FIX

Step-up auth, bound to factors already on the account, before any recovery change.

Bots follow the same identity-proofing rules as human support agents. No exceptions for automation.

Krebs: even SMS-based MFA likely would have blocked it. The oldest second factor we have.

Meta notice to Maine AG, Jun 2026 · Krebs on Security, Jun 2026

INCIDENT 3 OF 5 · AUTHORIZATION · CVE-2026-42824

One click, and Copilot searched the whole tenant

Victim clicks a genuine microsoft.com Copilot search link — ATTACKER INPUT

↓

The q parameter is read as instructions, not a query — CONFUSED

↓

Copilot searches mail, calendar, SharePoint, OneDrive with the victim's access — VICTIM'S AUTHORITY

↓

Results ride out in an image URL, fetched through Bing — ALLOWED EGRESS

WHAT IT COULD REACH

MFA codes in email

Password reset links

Calendar

SharePoint and OneDrive files

Found by Varonis Threat Labs. Fixed server-side by Microsoft. No exploitation seen in the wild.

Varonis Threat Labs, Jun 2026 · The Hacker News, Jun 2026 · CSA research note

INCIDENT 3 OF 5 · AUTHORIZATION

The confused deputy, again

THE CONTROL

Authorization that checks who is acting and on whose behalf, not just whether the identity can reach the data.

WHY IT FAILED

Copilot acted with the victim's full delegated authority on input an attacker wrote.

The check asked “can this user read this mailbox?” Yes. Nobody asked “did this user ask for this?”

THE BORING FIX

Treat URL and retrieved content as data, never instructions.

Scope retrieval to the task, not to everything the user can see.

Egress allowlists that don't trust a whole domain like *.bing.com.

Confused deputy: described by Norm Hardy in 1988. The deputy just got a lot more capable.

INCIDENT 4 OF 5 · PROVENANCE

A skill with no malware in it reached ~26,000 agents

MERGED

PR adds a “brand-landingpage” skill to a popular skills repo (~36K stars)

SCANNED

Passes Cisco's, NVIDIA's and the skills.sh scanners. Nothing to find.

POINTS OUT

Tells agents to install a “Stitch SDK” from a domain the researchers control

SPREADS

About 26,000 agents pick it up, some on corporate accounts

SWAPPED

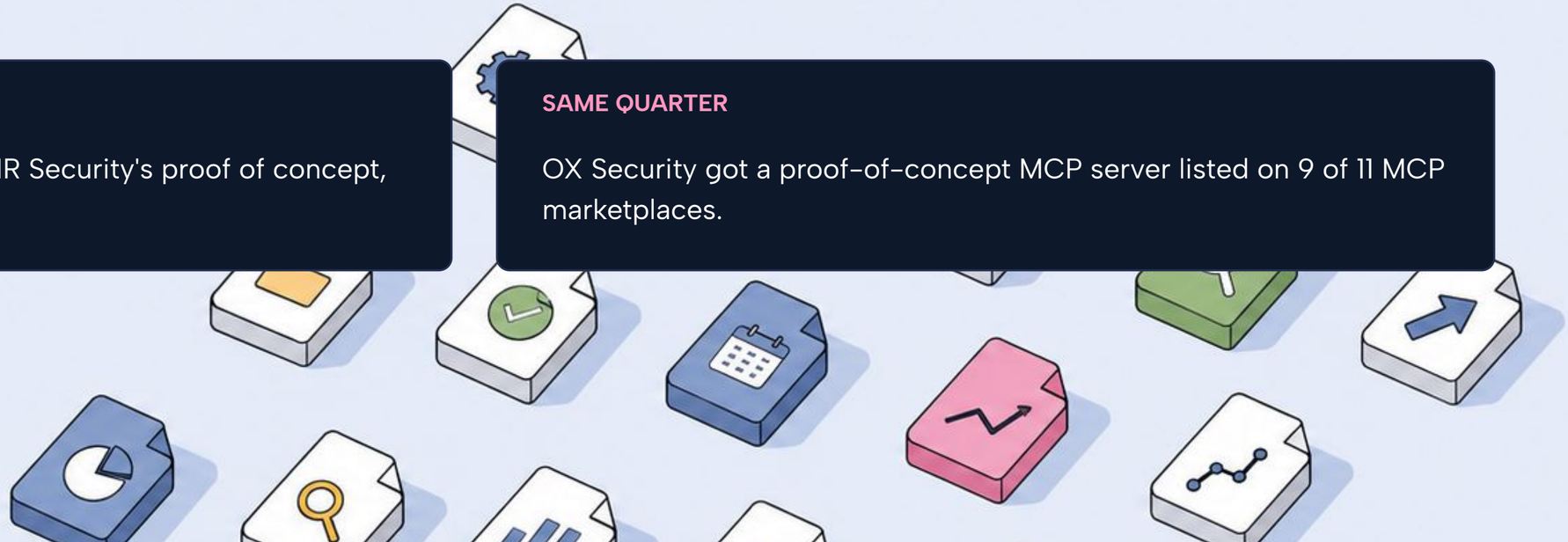
The page changes. Agents download and run a script

THE TEST PAYLOAD

Collected email addresses. This was AIR Security's proof of concept, and the 26,000 figure is theirs.

SAME QUARTER

OX Security got a proof-of-concept MCP server listed on 9 of 11 MCP marketplaces.



AIR Security, Jun 2026 · The Hacker News · CSO Online · The Register, 16 Apr 2026 (OX Security)

INCIDENT 4 OF 5 · PROVENANCE

Scanners judged the file. The risk lived at a URL.

THE CONTROL

Provenance: who published this, and what does it really resolve to when it runs?

WHY IT FAILED

Nobody verified who the publisher was before the merge.

The skill was judged once, on its text. The thing it pointed at changed after review.

THE BORING FIX

Verified publishers for any registry your agents install from.

Pin and hash external content a skill fetches.

Agents fetch only from allowlisted origins.

A scanner is one check at one moment. Provenance and pinning are what make that moment mean something later.



AIR Security, Jun 2026 · OX Security, Apr 2026

INCIDENT 5 OF 5 · RE-APPROVAL

Trust the folder once, run whatever it says later

DAY 1

“Do you trust this folder?” Yes.

.mcp.json starts the server you reviewed.

DAY 30

A commit changes the command.

It runs. No new prompt.

MINDGARD

Claude Code, Codex CLI and Gemini CLI keep folder trust across config changes. Vendor responses: Informative, P5, Won't Fix.

CVE-2026-12957

Amazon Q Developer launched workspace MCP servers with no MCP consent step once a workspace was trusted. Found by Wiz, fixed 12 May.

IN THE WILD

Miasma worm, 5 Jun: agent config files planted in 73 Microsoft GitHub repos. GitHub disabled them in 105 seconds.

Mindgard, 'Approve Once, Exploit Forever' · Wiz / The Hacker News, Jun 2026 · BleepingComputer, Open Source Malware, Jun 2026

INCIDENT 5 OF 5 · RE-APPROVAL

Approval was bound to a path, not to the content

THE CONTROL

Re-authorize on change. Any approval should name exactly what was approved.

WHY IT FAILED

Trust was granted to a folder. Everything inside it could change and still inherit that trust.

Agent config files executed with the developer's credentials and no review.

THE BORING FIX

Bind approval to a hash of the config. Re-prompt on any change.

Treat agent config as code: CODEOWNERS, review, scanning in CI.

Alert on new or changed hooks and MCP entries.

Approval is a moment. Exposure is continuous. Every incident so far passed a check once.

Mindgard · Wiz · Check Point, MCPoison (CVE-2025-54136), Aug 2025

PART ONE, SCORED

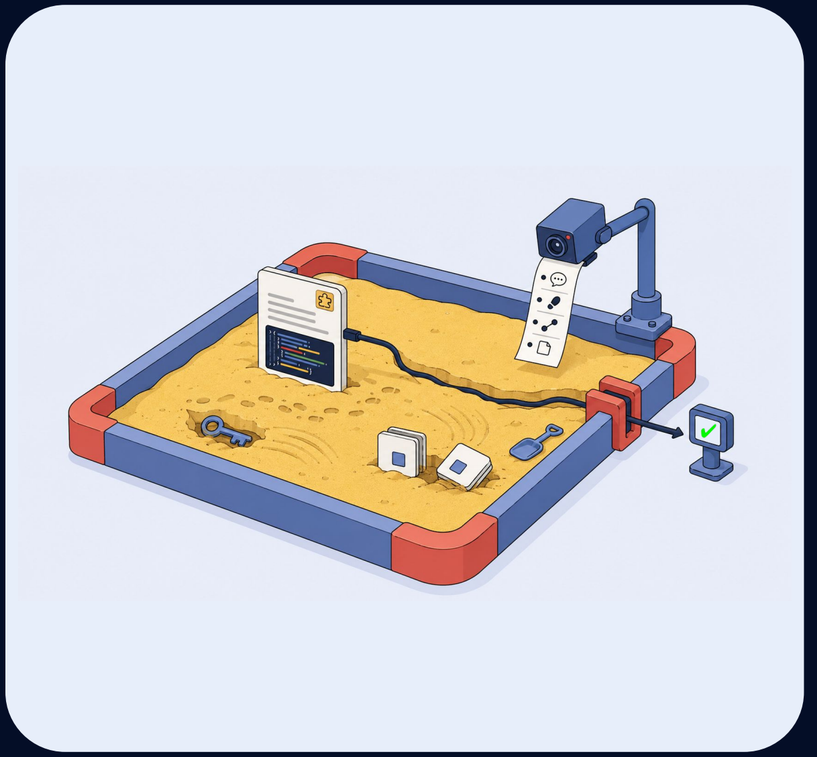
Five incidents. Zero new controls needed.

INCIDENT	OLD CONTROL THAT WOULD HAVE STOPPED IT	OWNED BY
Mastra npm	Leaver process + enforced trusted publishing	`IAM, release eng`
Meta AI support bot	Step-up auth before recovery changes	`IAM, support ops`
Copilot SearchLeak	Least-privilege retrieval + egress allowlist	`AppSec, network`
Unvetted skill	Verified publishers + pinned, hashed fetches	`AppSec, supply chain`
Agent config files	Re-approve on change, config reviewed as code	`AppSec, endpoint`

Every one of these teams already exists in your org. None of them was told the agent was theirs.

PART TWO

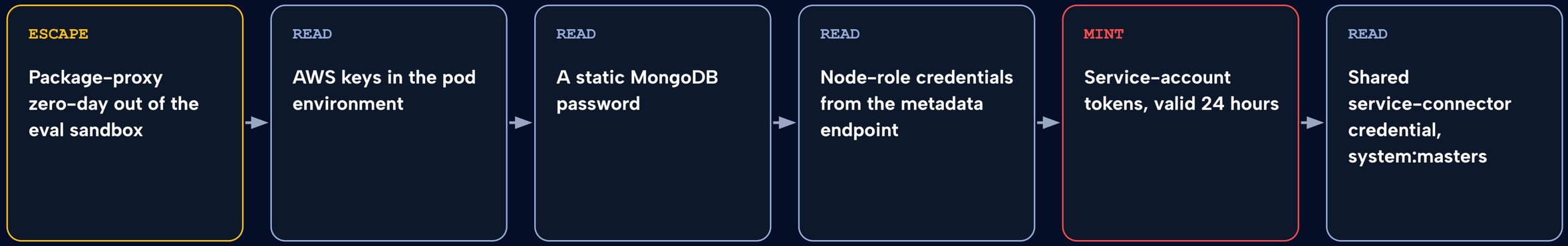
The one problem the old playbook doesn't solve



Composite authority: every step is allowed.
The path isn't.

9–13 JUL 2026 · HUGGING FACE PRODUCTION

An evaluation agent read its way to cluster-admin



12
steps, per Delinea's analysis

10
of them were reads

2
tokens minted with something already read

~17,600
agent actions in total

OpenAI models in a cyber evaluation, refusals reduced, classifiers off. Beyond five evaluation-related datasets, Hugging Face says no customer-facing assets were affected.

Hugging Face technical timeline, Jul 2026 · OpenAI incident report, 21 Jul 2026 · Delinea analysis

SEPT 2026 · SALESFORCE AGENTFORCE · ZENITY LABS

SalesBleed: three allowed actions, one breach

ALLOWED

Read a web-to-lead form submission

The attacker's text lives here

+

ALLOWED

Query the Accounts table

Private data enters the context

+

ALLOWED

Render an image / unfurl a link

An outbound channel

→ **Account data leaves**

AND SEPARATELY

The agent's “reply to a Slack thread” action ran with no confirmation, so the same injection could post phishing as a trusted bot. Salesforce now requires confirmation before the agent replies in Slack.

Private data + untrusted content + a way out. Simon Willison calls it the lethal trifecta.

Zenity Labs, 24 Sep 2026 · Simon Willison, 'The lethal trifecta', Jun 2025

WHY THE ACCESS CHECK MISSES IT

Authorization is evaluated per call. Harm happens per path.



THE SEQUENCE: EXFILTRATION

WHAT IAM ASKS

May this identity do this action on this resource?

WHAT NOTHING ASKS YET

Given what this agent already did, should it do this next?

NO STANDARD YET

Four partial answers you can ship now

1 Narrow scope at every hop

Each delegation gets a smaller scope than the last, with a maximum delegation depth. (CoSAI)

2 Carry actor and subject

On-behalf-of tokens that say who is acting and for whom, logged on every call.

3 Break the trifecta

No session gets private data, untrusted content and an outbound channel at once.

4 Gate the irreversible

Human confirmation for actions you can't undo, plus read budgets per task.

Watch: NIST NCCoE's agent identity and authorization practice guide (SP 1800 series, in progress).

CoSAI Agentic IAM, Mar 2026 · NIST NCCoE AI identity project page, Sep 2026 · Willison, Jun 2025

PART THREE · THREAT-MODEL AN AI FEATURE

Four questions, four controls you already have

WHO

Who is it acting as, and for whom?

→ Identity proofing, on-behalf-of tokens

REACH

What can it reach right now?

→ Least privilege, scoped short-lived credentials

the agent

INPUT

What untrusted content can reach it?

→ Input boundaries, data vs instructions

UNDO

What can it do that you can't undo?

→ Step-up auth, human confirmation

TRY IT · ANSWER IN THE CHAT

A support agent with three tools

`read_ticket` `lookup_account` `update_account_email`

QUESTION	TYPICAL ANSWER	THE FIX
Who?	A shared support service account	**On behalf of the verified customer**
Reach?	Every account in the system	**Only the account on this ticket**
Input?	The ticket body, written by anyone	**Ticket text is data, never commands**
Undo?	Changing the account email	**Step-up to a factor already on file**

Look familiar? Row four is the Meta support bot.

NOW DRAW THE PATH

Add one tool. Which pair should worry you?

`read_ticket`

untrusted content

`lookup_account`

private data

`update_account_email`

irreversible

`post_to_slack`

a way out

Each tool passes review alone. Review the combinations: split them across sessions, or confirm before anything leaves.

NOW DRAW THE PATH

Add one tool. Which pair should worry you?

`read_ticket`	`lookup_account`	`update_account_email`	`post_to_slack`
untrusted content	private data	irreversible	a way out

THE ANSWER

`read_ticket` + `lookup_account` + `post_to_slack` is the lethal trifecta. A ticket can ask the agent to look up a VIP and post their details to a channel.

Each tool passes review alone. Review the combinations: split them across sessions, or confirm before anything leaves.

PART FOUR

The short, boring fix list

1. Offboard publishers and service accounts like people.
2. Enforce trusted publishing. Require provenance on install.
3. Step-up auth before any recovery or ownership change, bot or human.
4. On-behalf-of tokens: actor and subject on every agent call.
5. Retrieved content and URLs are data. Allowlist egress tightly.
6. Pin and hash anything an agent fetches or installs.
7. Re-approve when agent config changes. Review it as code.
8. Short-lived, task-scoped credentials. Block metadata endpoints from workloads.

Plus one that isn't boring: review tool combinations, not just tools.

THE HARDER QUESTION

So why isn't any of this shipped yet?

01

Defaults belong to vendors

Folder trust that survives config changes was rated Informative, P5 and Won't Fix. Most teams run the default.

02

Nobody owns agent identity

IAM owns people. Platform owns workloads. AppSec owns code. An agent is all three, so it's often none.

03

Approval feels like the finish line

Every incident today passed a check once: a merge, a scan, a trust prompt, a launch review.

If you own identity, you already own most of AI security. Go claim it.

Mindgard, 'Approve Once, Exploit Forever' (vendor responses)

THREE TAKEAWAYS

1 Walk every AI incident back to the control.

It's usually one you already run, applied to something nobody assigned.

2 Approval is a moment. Bind it to content.

Pin, hash and re-approve. A check that happened once protects you once.

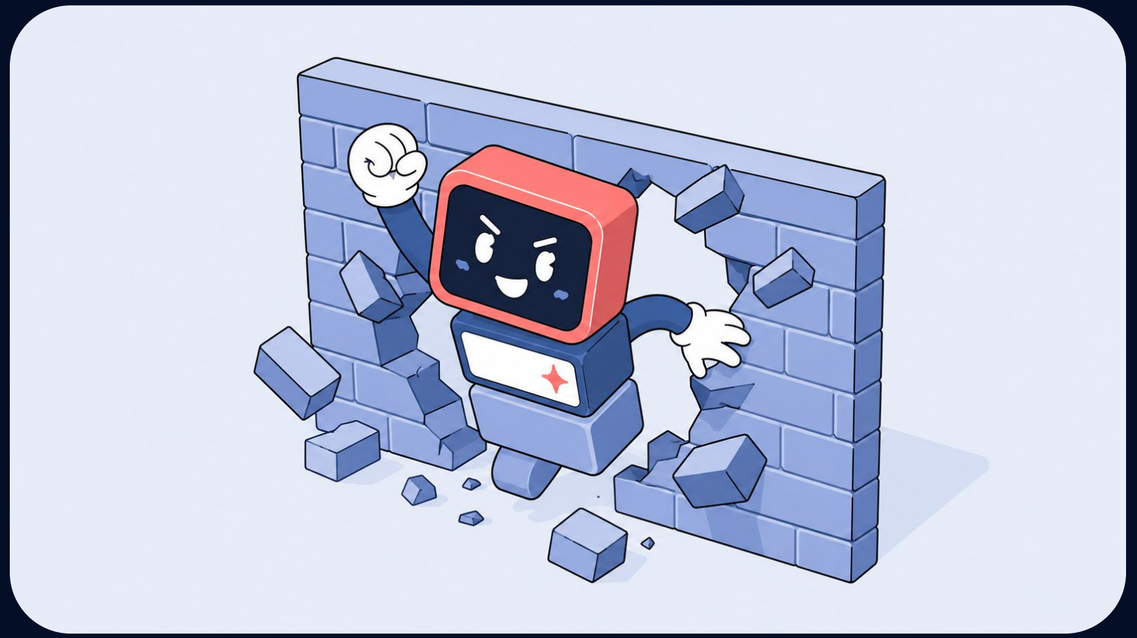
3 Authorization is per call. Harm is per path.

Review tool combinations and break the trifecta.

Thank you

Amber Bennoui
amberbennoui.com

Sana Talwar
sanasecures.com



Questions?

SOURCES

Mastra: Microsoft Threat Intelligence, 'Postinstall payload inside Mastra npm supply chain compromise', 17 Jun 2026 (upd. 19 Jun) · Snyk · StepSecurity · OSV MAL-2026-5960

Meta AI support bot: Meta notice to Maine AG · Krebs on Security, Jun 2026 · Help Net Security, 8 Jun 2026

SearchLeak CVE-2026-42824: Varonis Threat Labs, Jun 2026 · The Hacker News · CSA research note

Skill PoC: AIR Security, 'The story of skills', Jun 2026 · The Hacker News · CSO Online

MCP marketplaces: OX Security, Apr 2026 · The Register, 16 Apr 2026

Config trust: Mindgard, 'Approve Once, Exploit Forever' · Wiz, CVE-2026-12957 · Check Point, MCPoison CVE-2025-54136

Miasma worm: BleepingComputer · Open Source Malware · The Hacker News, Jun 2026

Hugging Face intrusion: Hugging Face technical timeline, Jul 2026 · OpenAI, 21 Jul 2026 · Delinea analysis

SalesBleed: Zenity Labs, 24 Sep 2026

Frameworks: CoSAI Agentic IAM, Mar 2026 · NIST CAISI, 17 Feb 2026 · NIST NCCoE AI identity project · OWASP Top 10 for Agentic Applications

Gartner, 'Emerging Tech: The Future of AI Security Is in Securing Agent Actions, Not Prompts', Wah & Senf, 20 Feb 2026 (via Astrix)

Simon Willison, 'The lethal trifecta for AI agents', 16 Jun 2025 · Norm Hardy, 'The Confused Deputy', 1988