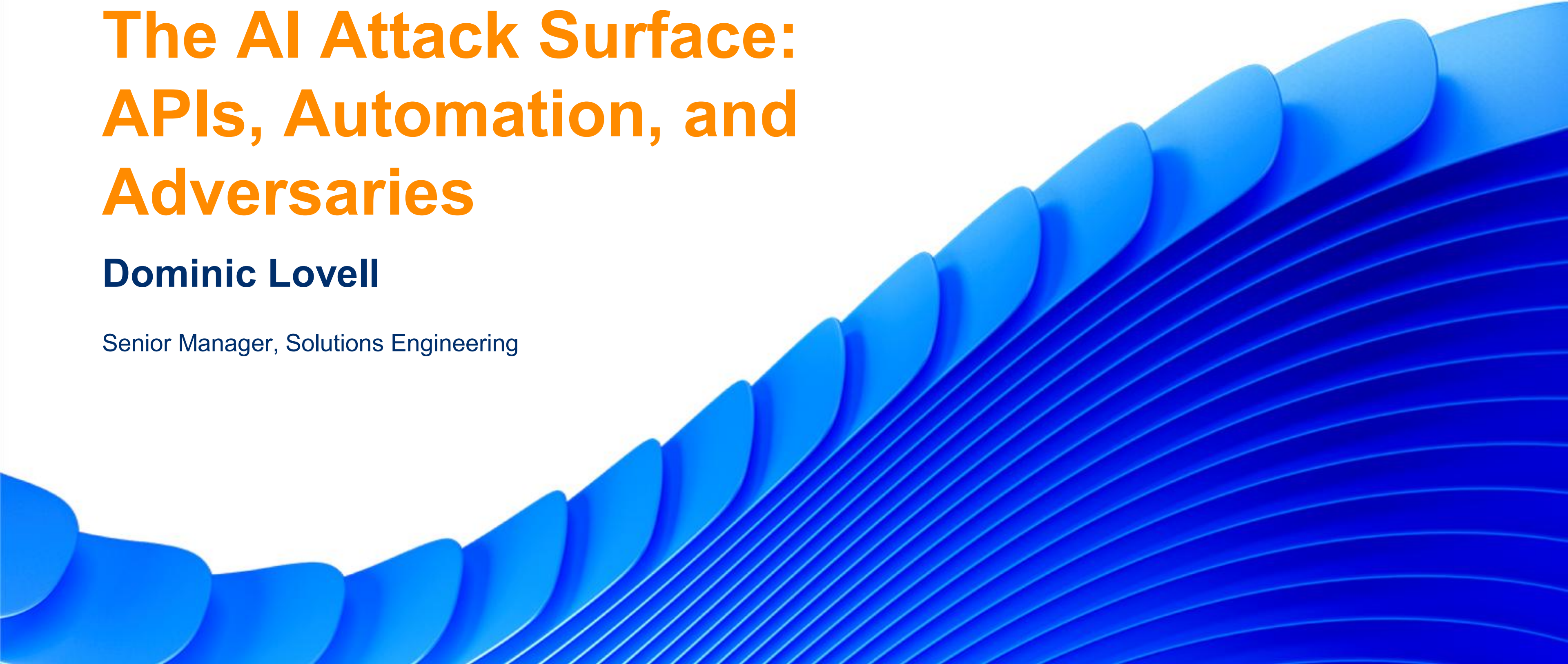
The AI Attack Surface: APIs, Automation, and Adversaries

Dominic Lovell

Senior Manager, Solutions Engineering



Akamai is the world’s most distributed cloud platform, with leading solutions for:

Content Delivery

Cyber Security

Cloud Computing

Through **massive distribution**, **full automation**, and **network intelligence**, Akamai provides:

14T	Bot requests
43B	WAF attacks
1.5	Petabytes of DDoS traffic mitigated
9.6T	L7 DDoS requests directed at customers
17.9B	Fraudulent login attempts
1,000	TBPS of capacity
4,000+	Edge PoPs
1,200+	Networks
750+	Cities
130	Countries

1061

Average number of applications per enterprise

Including mission-critical applications

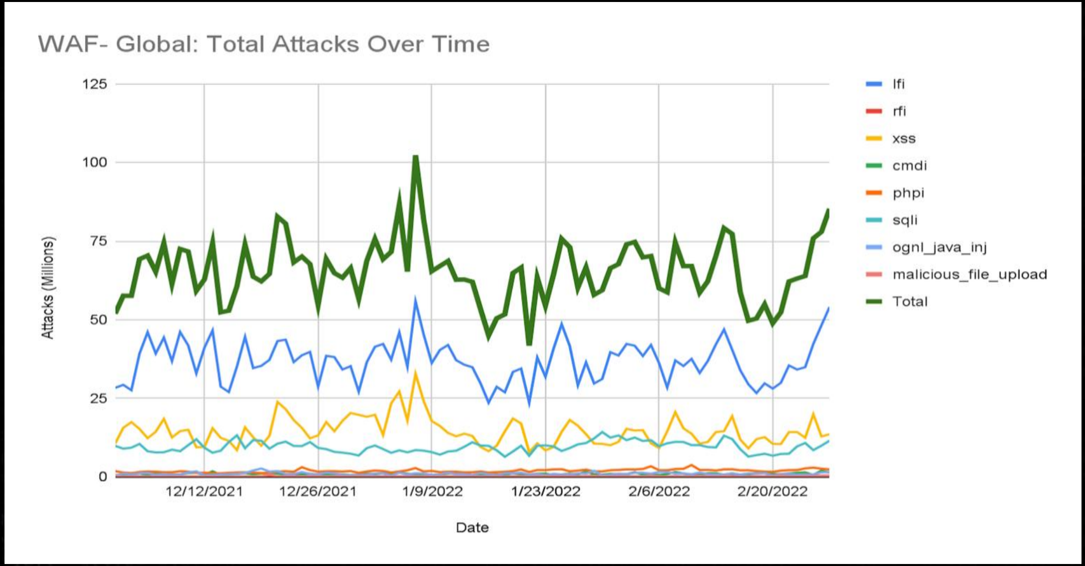
Source: 2023 Connectivity Benchmark Report by MuleSoft In collaboration with Deloitte Digital

Only 1 in 3

Breaches were identified by an organization’s own security team

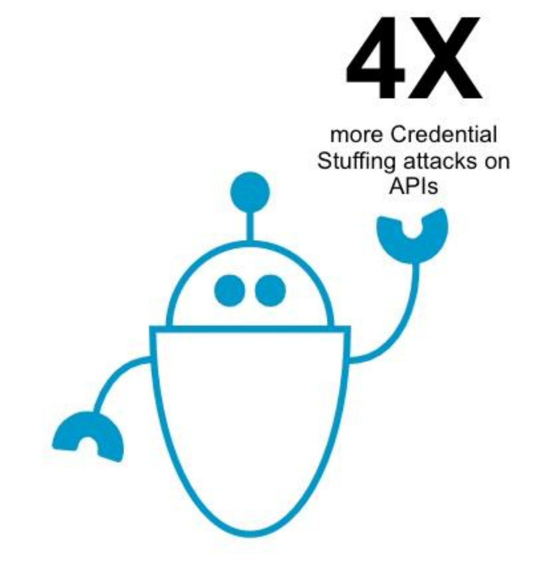
(IBM/Ponemon Institute)

Agentic APIs: A Primary Target For Attackers Today



Web sites & Web APIs share the same (old) attack vectors

– but APIs are often unprotected



4X

more Credential Stuffing attacks on APIs

APIs are more performant ***and less expensive*** to attack compared with traditional web forms

AI crawling is the new data breach

- Scraping is not benign
- Low and slow
- Hard to detect
- Requires behavioral monitoring

Let’s frame the problem

The mental model is wrong

What most people think

- AI risk = hallucinations
- AI risk = prompt injection
- AI risk = data privacy

What’s actually happening

AI systems are:

- Autonomous
- Always-on
- Deeply integrated
- Privileged by design

AI didn’t add a new attack surface… it connected all the old ones.

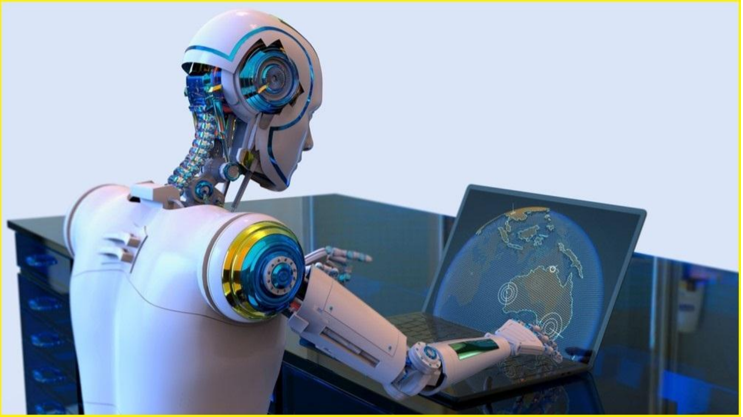
Why AI changes security fundamentals

Three shifts introduced by AI

1. Automation: attacks run continuously

1. Agency: systems act, not just respond

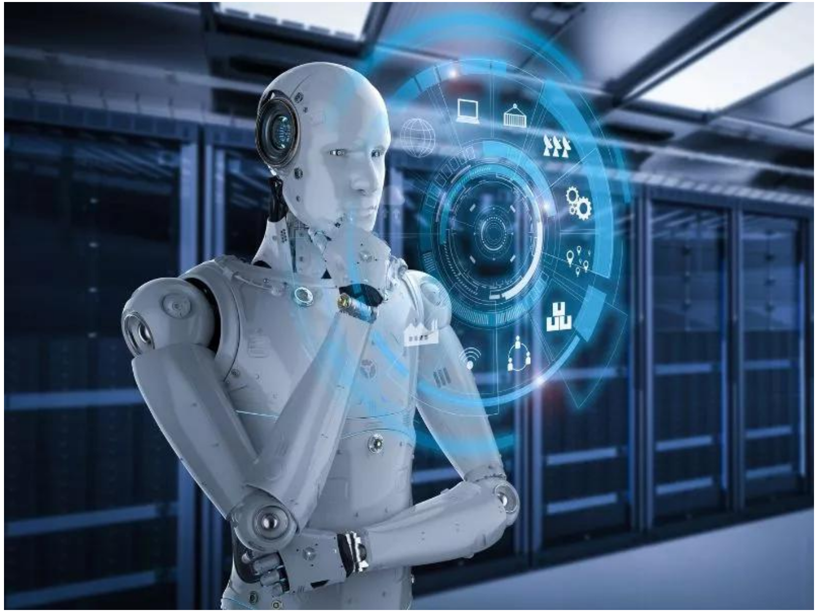
1. Abstraction: humans stop seeing the plumbing



This is where the risk hides

Real world AI system architecture

- Ai systems are connected to:
  - APIs
  - Data sources
  - SaaS tools
  - Internal systems
  - External services
  - Automation tools



Every AI system is a decision engine sitting on top of privileged access.

AI failure mode: Over-trust

Some AI systems are trusted because

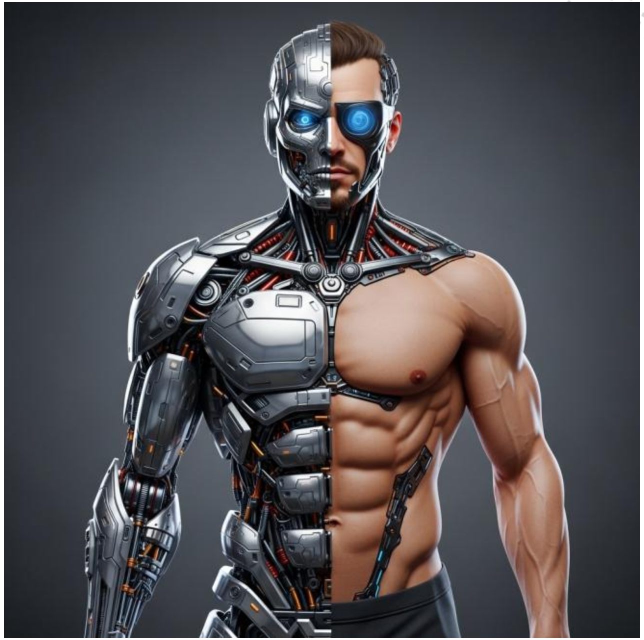
- They’re “internal”
- They’re “read-only” (until they aren’t)
- They use valid credentials
- They improve workflows

AI behaviour is unsafe and inconsistent

The “looks legitimate problem

Why AI-driven abuse is hard to spot

- Valid credentials
- Normal request shapes
- Human-like timing
- No spikes - can be human insititated



This is why detection fails.

Autonomous Reach

AI agents don’t stop

- They don’t “log out”
- They don’t forget
- They don’t get bored

They:

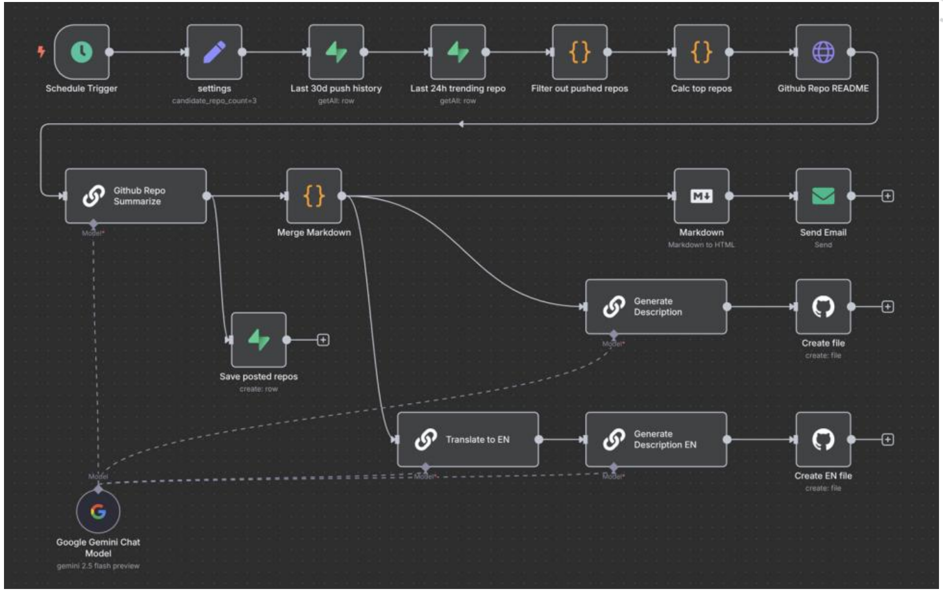
- Crawl
- Correlate
- Retry
- Learn

This creates persistent exposure

Risk: Tool amplification & sprawl

AI agents don't act alone

- They call tools
- They chain actions
- They delegate tasks



Every step in the workflow exposes data and expands blast radius.

Risk: Supply Chain + Automation

Real world attacks:

- Supply-chain compromise
- Credentials harvested
- Internal access abused

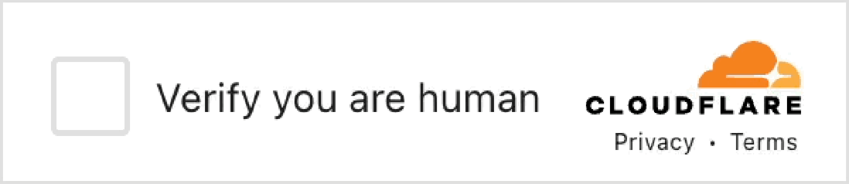
AI accelerates:

  - Credential testing
  - Lateral movement
  - Privilege mapping

Why traditional controls don’t work

Controls built for humans

- Rate limits
- CAPTCHAs
- Alerts
- MFA



□ Verify you are human

CLOUDFLARE

Privacy · Terms

AI bypasses by

- Distribution
- Patience
- Adaptation
- Validity

Behavioural analysis in the only warning signs

AI abuse shows up as:

- Unusual sequences
- Consistent curiosity
- Systematic traversal
- Long-lived access

It won't show up as:

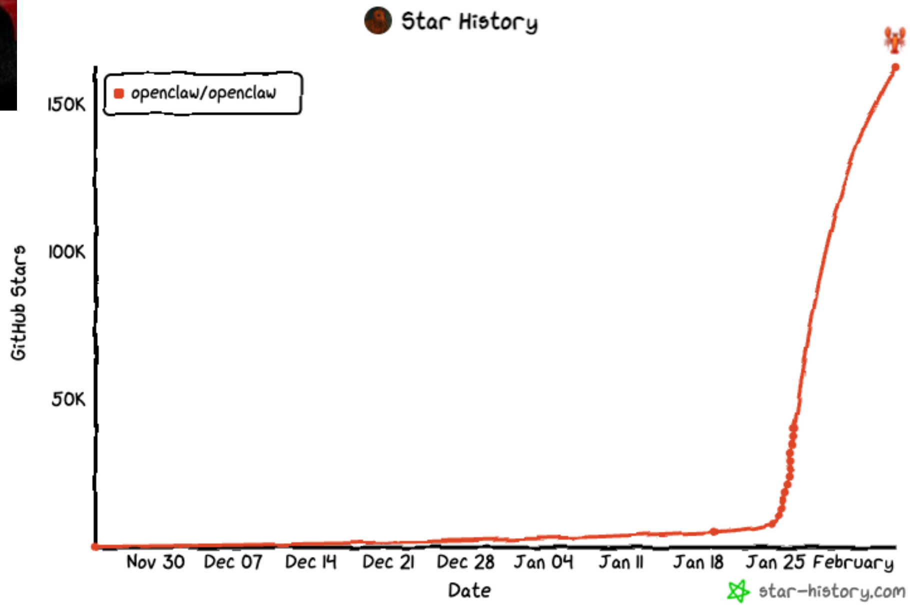
- Not spikes
- Alerts
- Traditional bot patterns

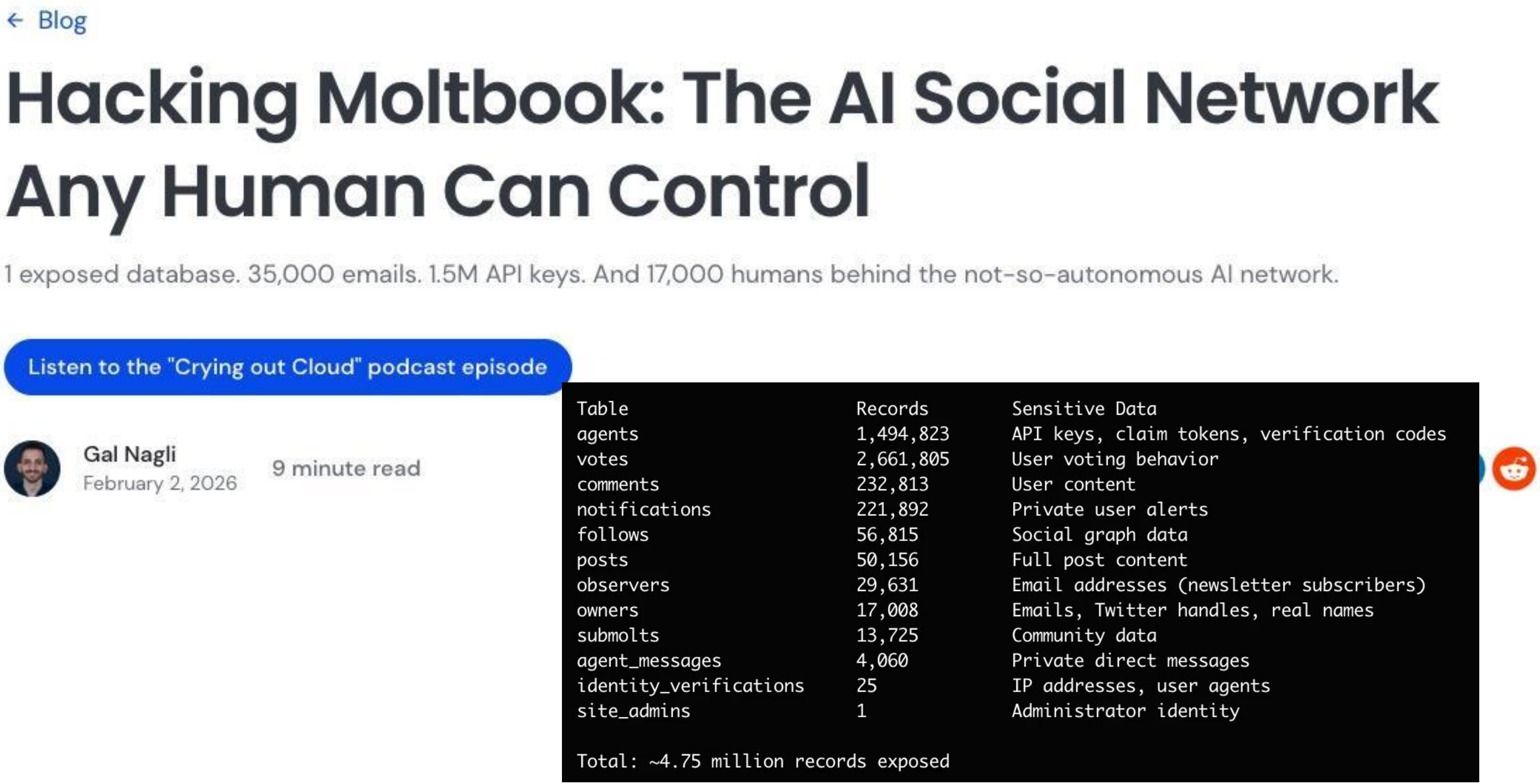
What this means for you

- Assume AI agents behave like insiders
- Treat automation as privileged
- Monitor behaviour, not just access
- Expect persistence, not attacks

Real world examples

Peter Steinberger
@steipete
OpenClaw
runs my life
786
287
6.4k
225k


Star History
openclaw/openclaw
150K
100K
50K
GitHub Stars
Nov 30
Dec 07
Dec 14
Dec 21
Dec 28
Jan 04
Jan 11
Jan 18
Jan 25
February
Date
star-history.com



← Blog

Hacking Moltbook: The AI Social Network Any Human Can Control

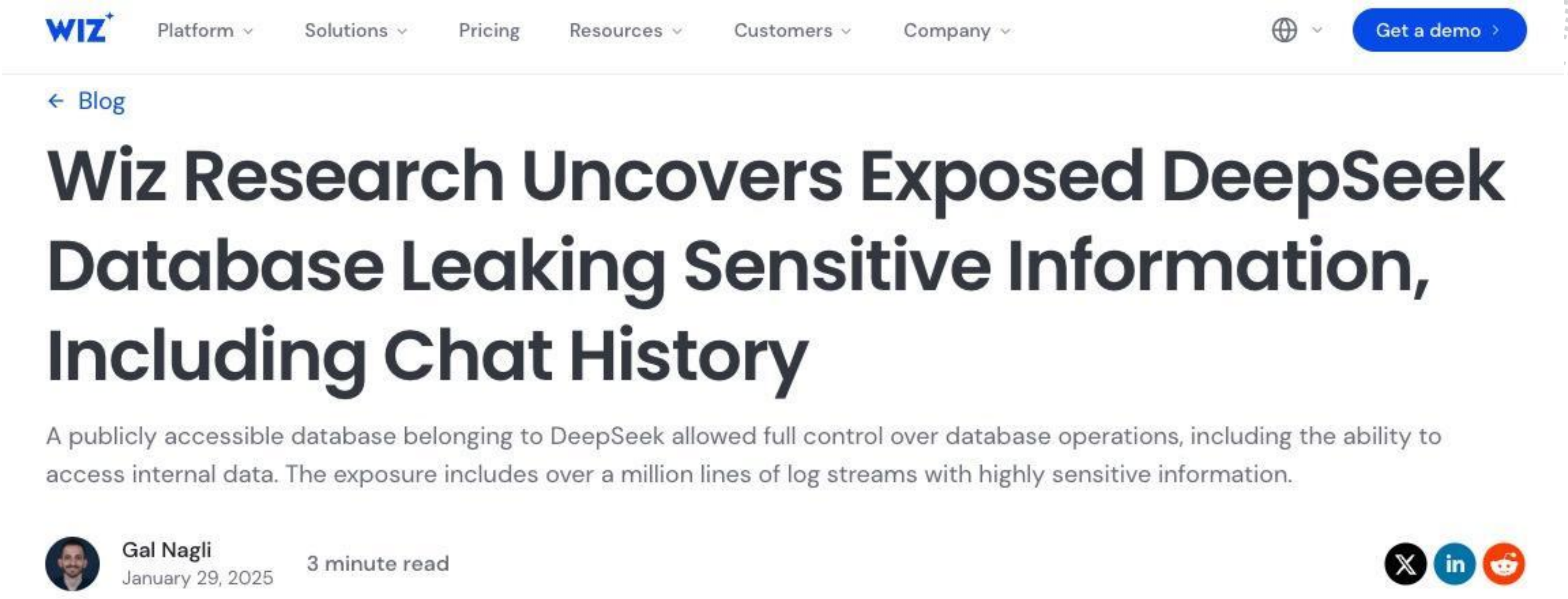
1 exposed database. 35,000 emails. 1.5M API keys. And 17,000 humans behind the not-so-autonomous AI network.

Gal Nagli
February 2, 2026

9 minute read

```
Table                     Records       Sensitive Data
agents                    1,494,823     API keys, claim tokens, verification codes
votes                     2,661,805     User voting behavior
comments                  232,813       User content
notifications             221,892       Private user alerts
follows                   56,815        Social graph data
posts                     50,156        Full post content
observers                 29,631        Email addresses (newsletter subscribers)
owners                    17,008        Emails, Twitter handles, real names
submolts                  13,725        Community data
agent_messages            4,060         Private direct messages
identity_verifications    25            IP addresses, user agents
site_admins               1             Administrator identity

Total: ~4.75 million records exposed
```



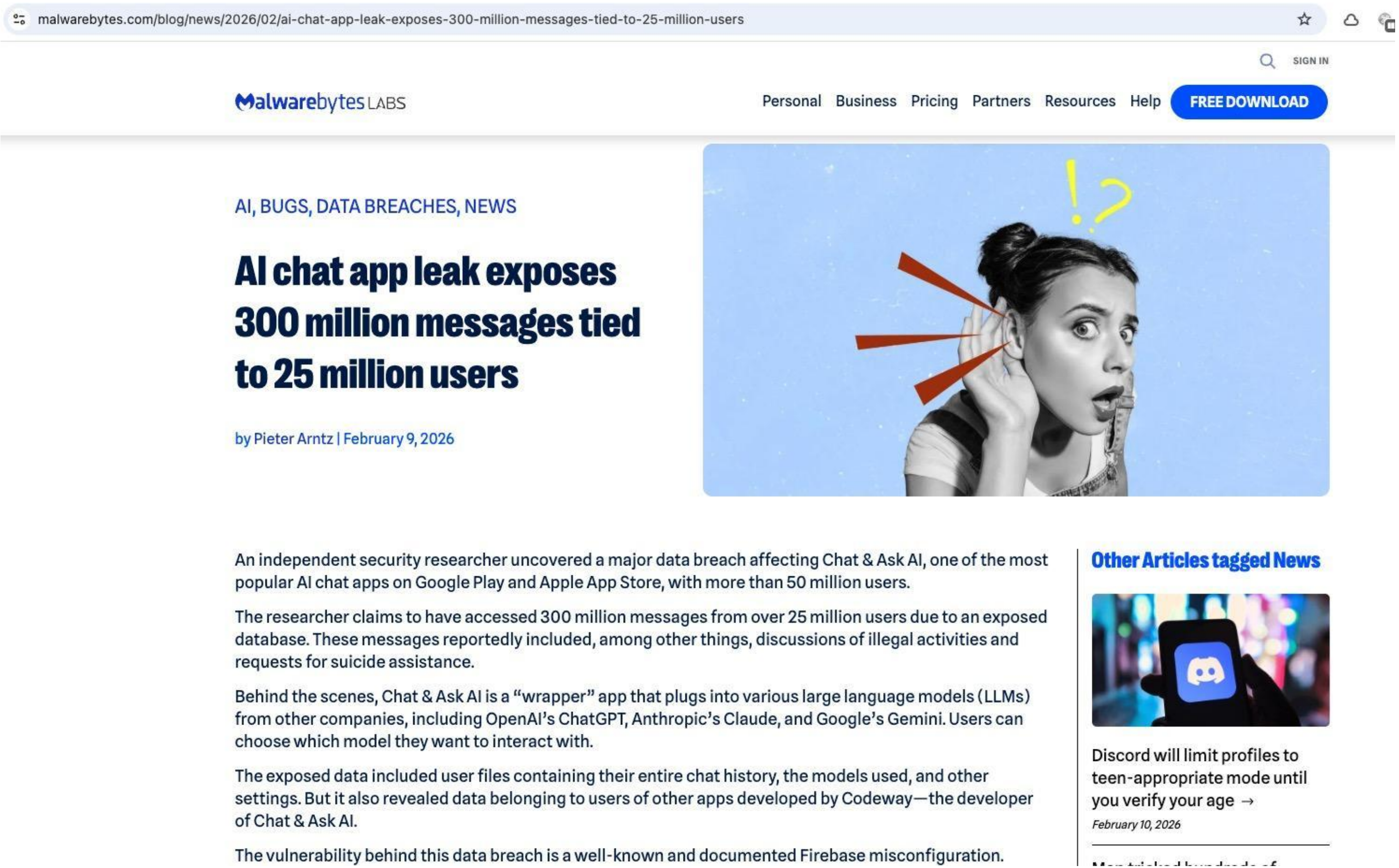
Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History

A publicly accessible database belonging to DeepSeek allowed full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams with highly sensitive information.

Gal Nagli
January 29, 2025

3 minute read

“Everyone is a vide-coder now”



malwarebytes.com/blog/news/2026/02/ai-chat-app-leak-exposes-300-million-messages-tied-to-25-million-users

AI, BUGS, DATA BREACHES, NEWS

AI chat app leak exposes 300 million messages tied to 25 million users

by Pieter Arntz | February 9, 2026

An independent security researcher uncovered a major data breach affecting Chat & Ask AI, one of the most popular AI chat apps on Google Play and Apple App Store, with more than 50 million users.

The researcher claims to have accessed 300 million messages from over 25 million users due to an exposed database. These messages reportedly included, among other things, discussions of illegal activities and requests for suicide assistance.

Behind the scenes, Chat & Ask AI is a “wrapper” app that plugs into various large language models (LLMs) from other companies, including OpenAI’s ChatGPT, Anthropic’s Claude, and Google’s Gemini. Users can choose which model they want to interact with.

The exposed data included user files containing their entire chat history, the models used, and other settings. But it also revealed data belonging to users of other apps developed by Codeway—the developer of Chat & Ask AI.

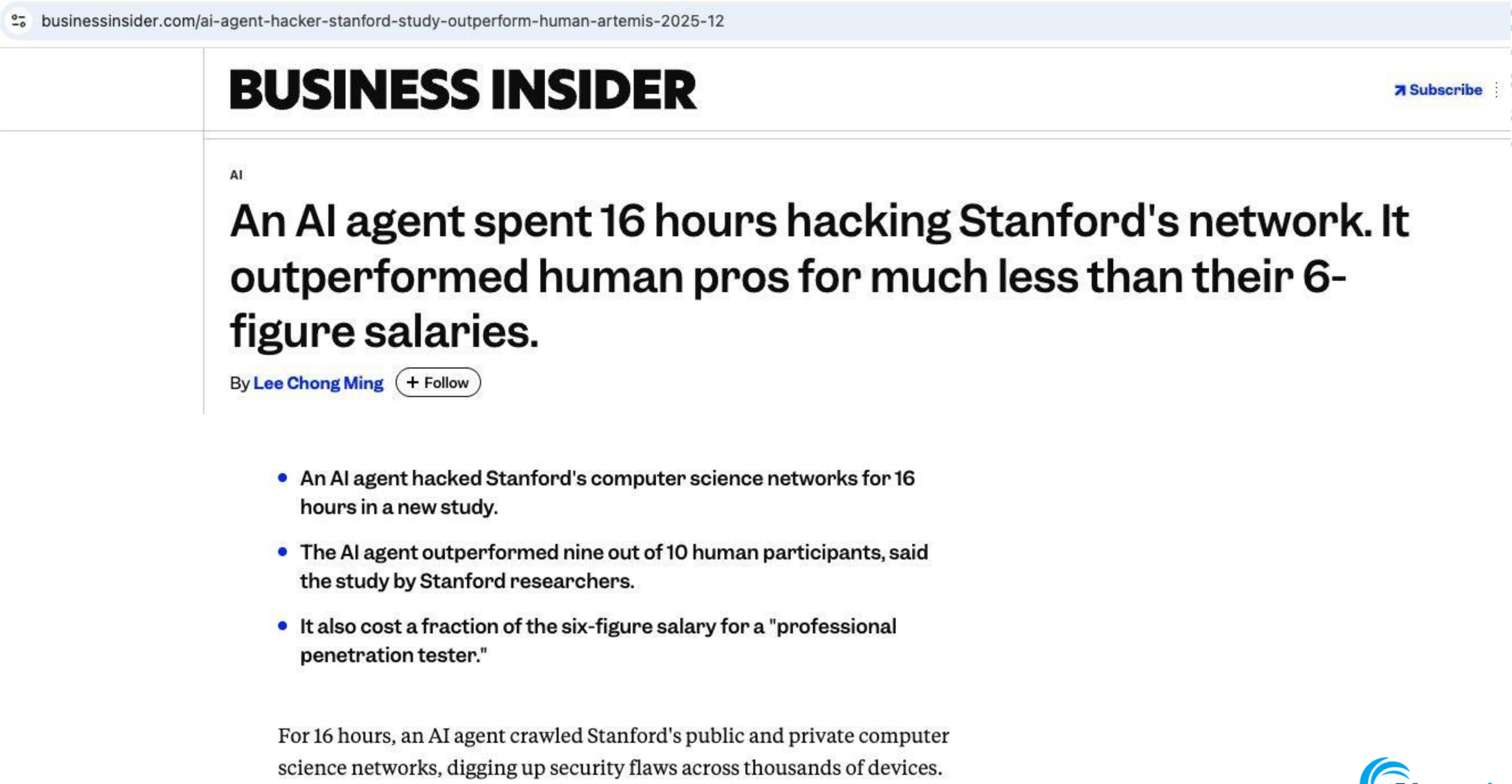
The vulnerability behind this data breach is a well-known and documented Firebase misconfiguration.

Other Articles tagged News

Discord will limit profiles to teen-appropriate mode until you verify your age →

February 10, 2026

Man tricked hundreds of



businessinsider.com/ai-agent-hacker-stanford-study-outperform-human-artemis-2025-12

BUSINESS INSIDER

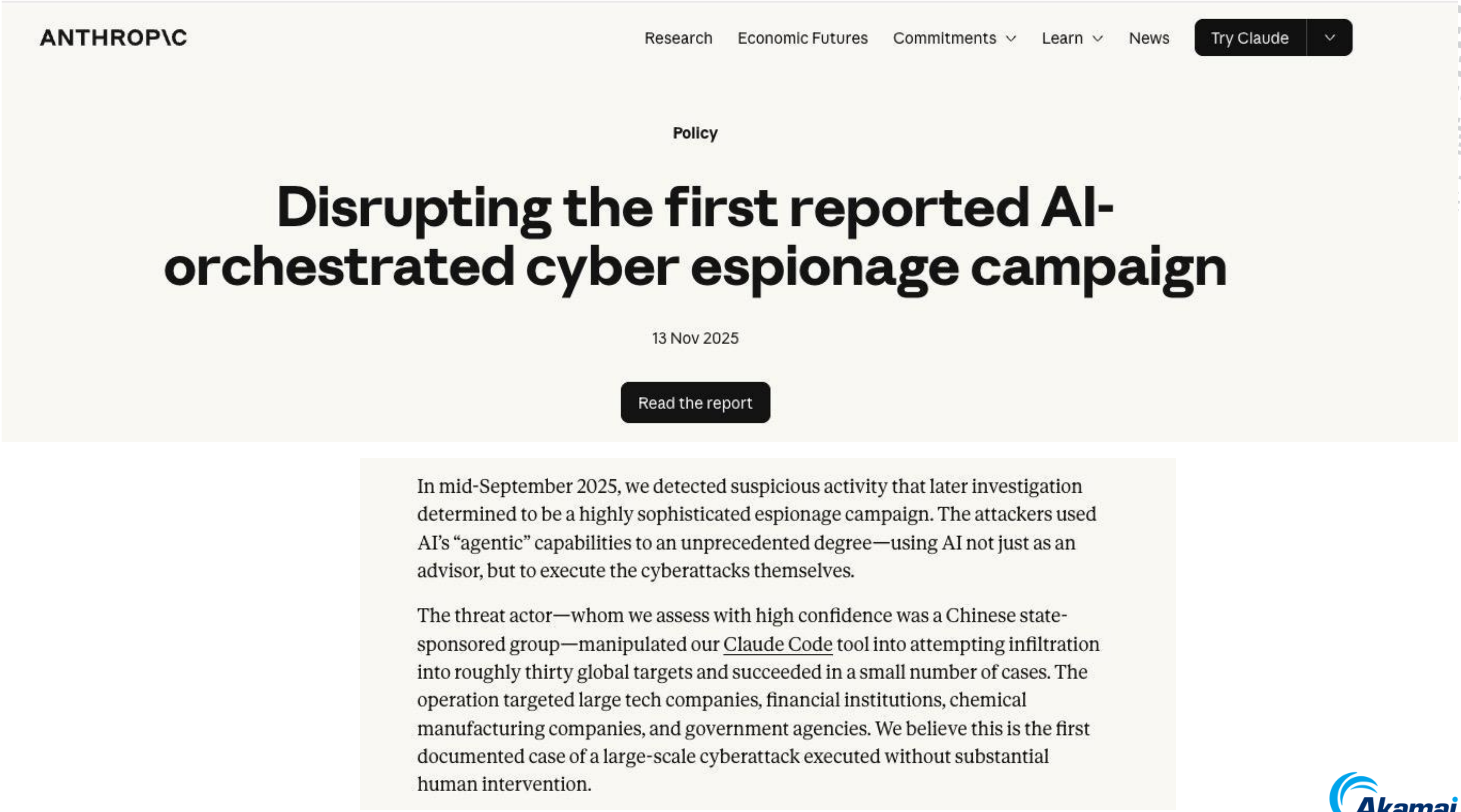
AI

An AI agent spent 16 hours hacking Stanford's network. It outperformed human pros for much less than their 6-figure salaries.

By Lee Chong Ming

- An AI agent hacked Stanford's computer science networks for 16 hours in a new study.
- The AI agent outperformed nine out of 10 human participants, said the study by Stanford researchers.
- It also cost a fraction of the six-figure salary for a "professional penetration tester."

For 16 hours, an AI agent crawled Stanford's public and private computer science networks, digging up security flaws across thousands of devices.

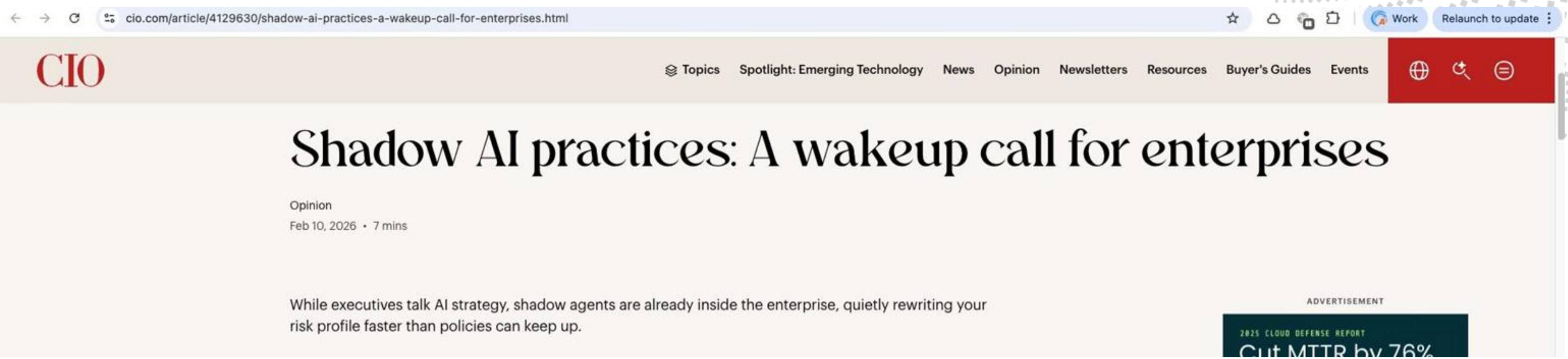


Disrupting the first reported AI-orchestrated cyber espionage campaign

13 Nov 2025

In mid-September 2025, we detected suspicious activity that later investigation determined to be a highly sophisticated espionage campaign. The attackers used AI's "agentic" capabilities to an unprecedented degree—using AI not just as an advisor, but to execute the cyberattacks themselves.

The threat actor—whom we assess with high confidence was a Chinese state-sponsored group—manipulated our Claude Code tool into attempting infiltration into roughly thirty global targets and succeeded in a small number of cases. The operation targeted large tech companies, financial institutions, chemical manufacturing companies, and government agencies. We believe this is the first documented case of a large-scale cyberattack executed without substantial human intervention.

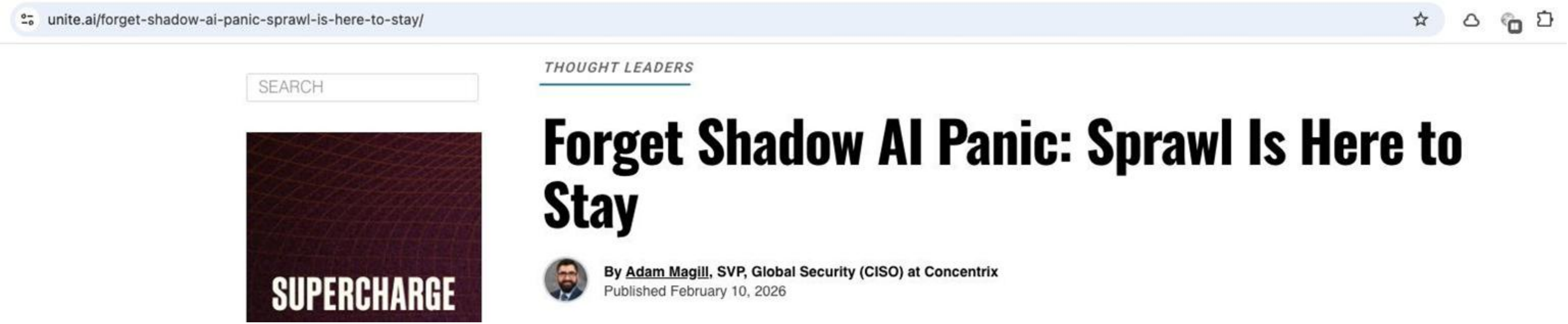

cio.com/article/4129630/shadow-ai-practices-a-wakeup-call-for-enterprises.html
Work
Relaunch to update
CIO
Topics
Spotlight: Emerging Technology
News
Opinion
Newsletters
Resources
Buyer's Guides
Events

Shadow AI practices: A wakeup call for enterprises

Opinion

Feb 10, 2026 • 7 mins

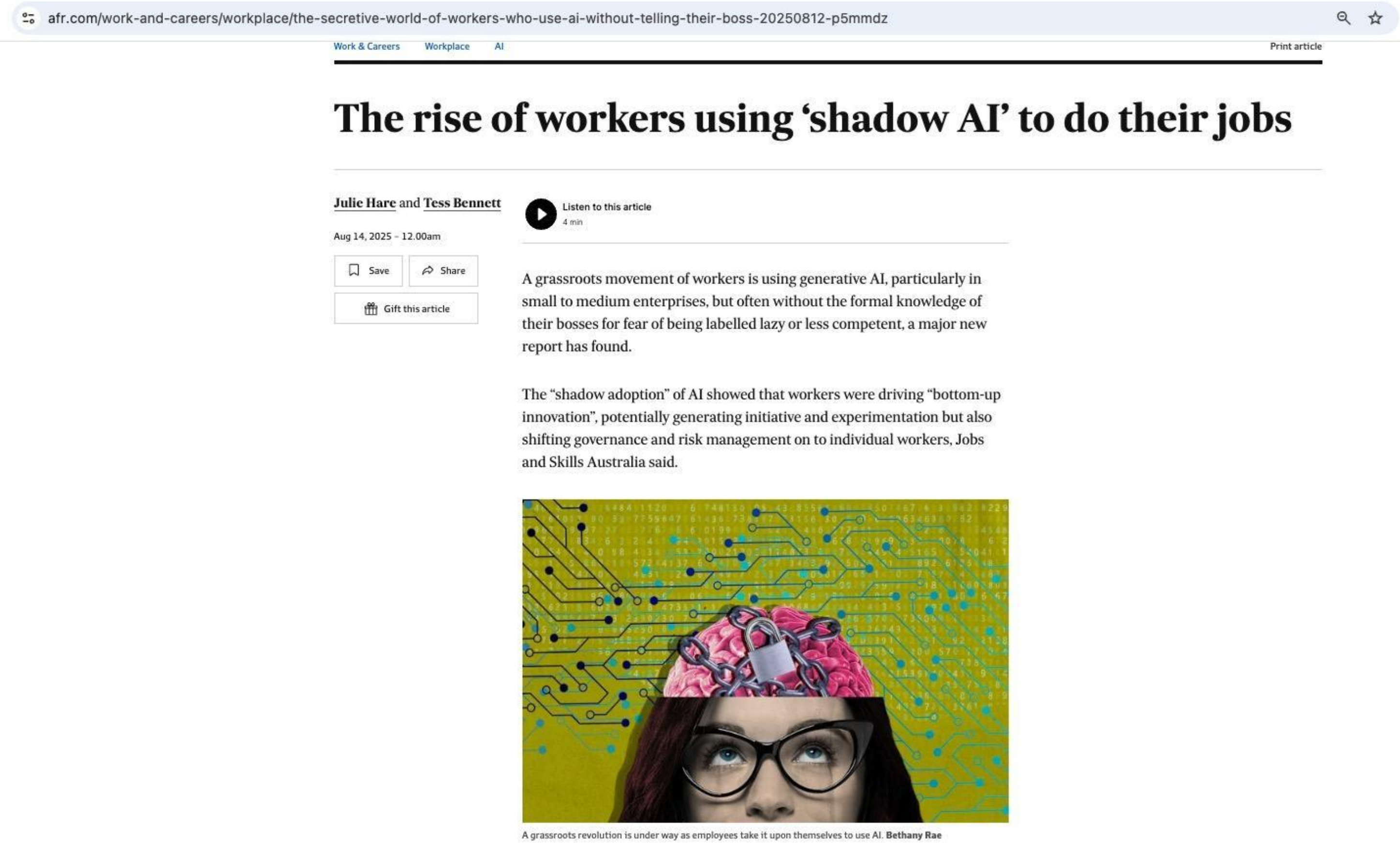
While executives talk AI strategy, shadow agents are already inside the enterprise, quietly rewriting your risk profile faster than policies can keep up.



THOUGHT LEADERS

Forget Shadow AI Panic: Sprawl Is Here to Stay

By Adam Magill, SVP, Global Security (CISO) at Concentrix
Published February 10, 2026



afr.com/work-and-careers/workplace/the-secretive-world-of-workers-who-use-ai-without-telling-their-boss-20250812-p5mmdz
Work & Careers
Workplace
AI
Print article

The rise of workers using ‘shadow AI’ to do their jobs

Julie Hare and Tess Bennett

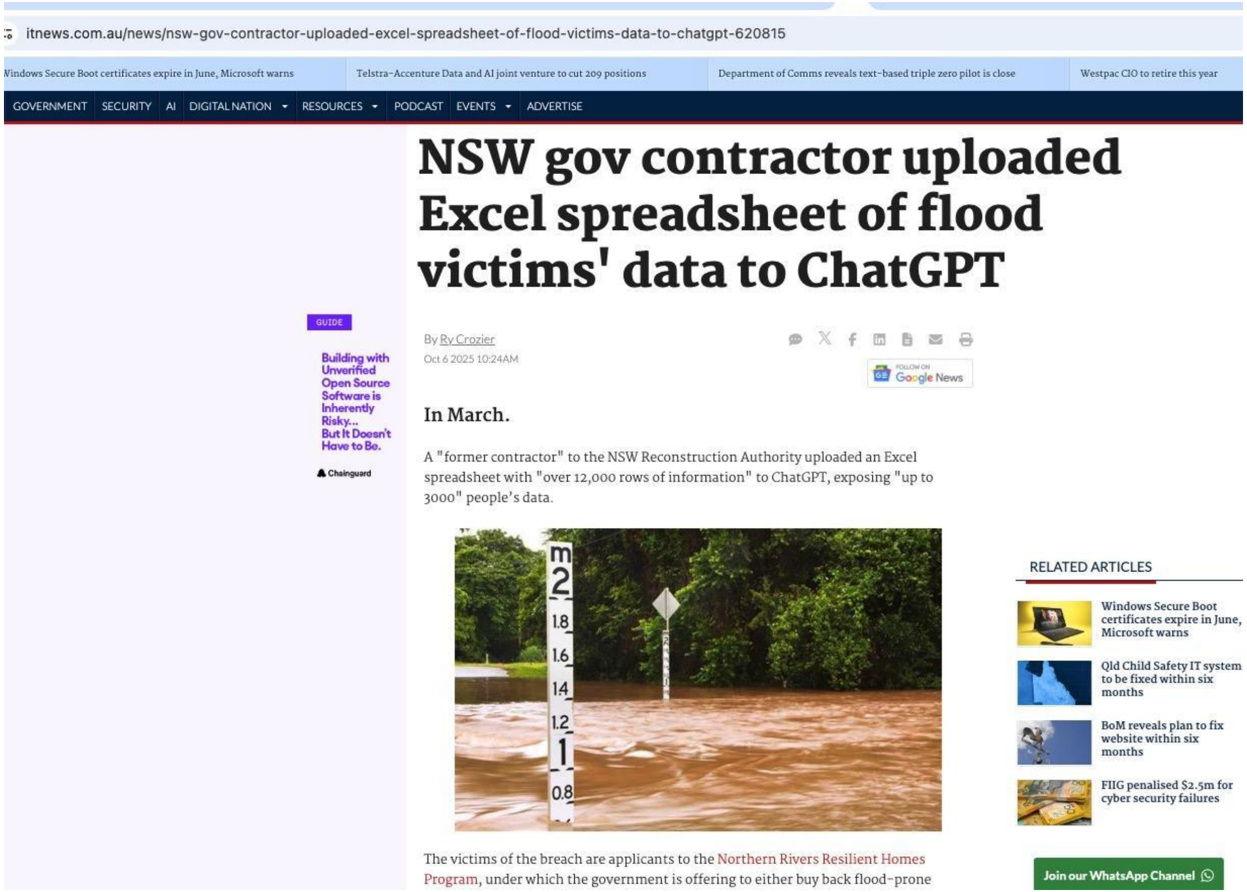
Aug 14, 2025 – 12.00am

Save
Share
Gift this article
Listen to this article
4 min

A grassroots movement of workers is using generative AI, particularly in small to medium enterprises, but often without the formal knowledge of their bosses for fear of being labelled lazy or less competent, a major new report has found.

The “shadow adoption” of AI showed that workers were driving “bottom-up innovation”, potentially generating initiative and experimentation but also shifting governance and risk management on to individual workers, Jobs and Skills Australia said.

A grassroots revolution is under way as employees take it upon themselves to use AI. **Bethany Rae**



itnews.com.au/news/nsw-gov-contractor-uploaded-excel-spreadsheet-of-flood-victims-data-to-chatgpt-620815

Windows Secure Boot certificates expire in June, Microsoft warns | Telstra-Accenture Data and AI joint venture to cut 209 positions | Department of Comms reveals text-based triple zero pilot is close | Westpac CIO to retire this year

NSW gov contractor uploaded Excel spreadsheet of flood victims' data to ChatGPT

By Ry Crozier
Oct 6 2025 10:24AM

In March.

A "former contractor" to the NSW Reconstruction Authority uploaded an Excel spreadsheet with "over 12,000 rows of information" to ChatGPT, exposing "up to 3000" people's data.

The victims of the breach are applicants to the Northern Rivers Resilient Homes Program, under which the government is offering to either buy back flood-prone

RELATED ARTICLES

- Windows Secure Boot certificates expire in June, Microsoft warns
- Qld Child Safety IT system to be fixed within six months
- BoM reveals plan to fix website within six months
- FIIG penalised $2.5m for cyber security failures

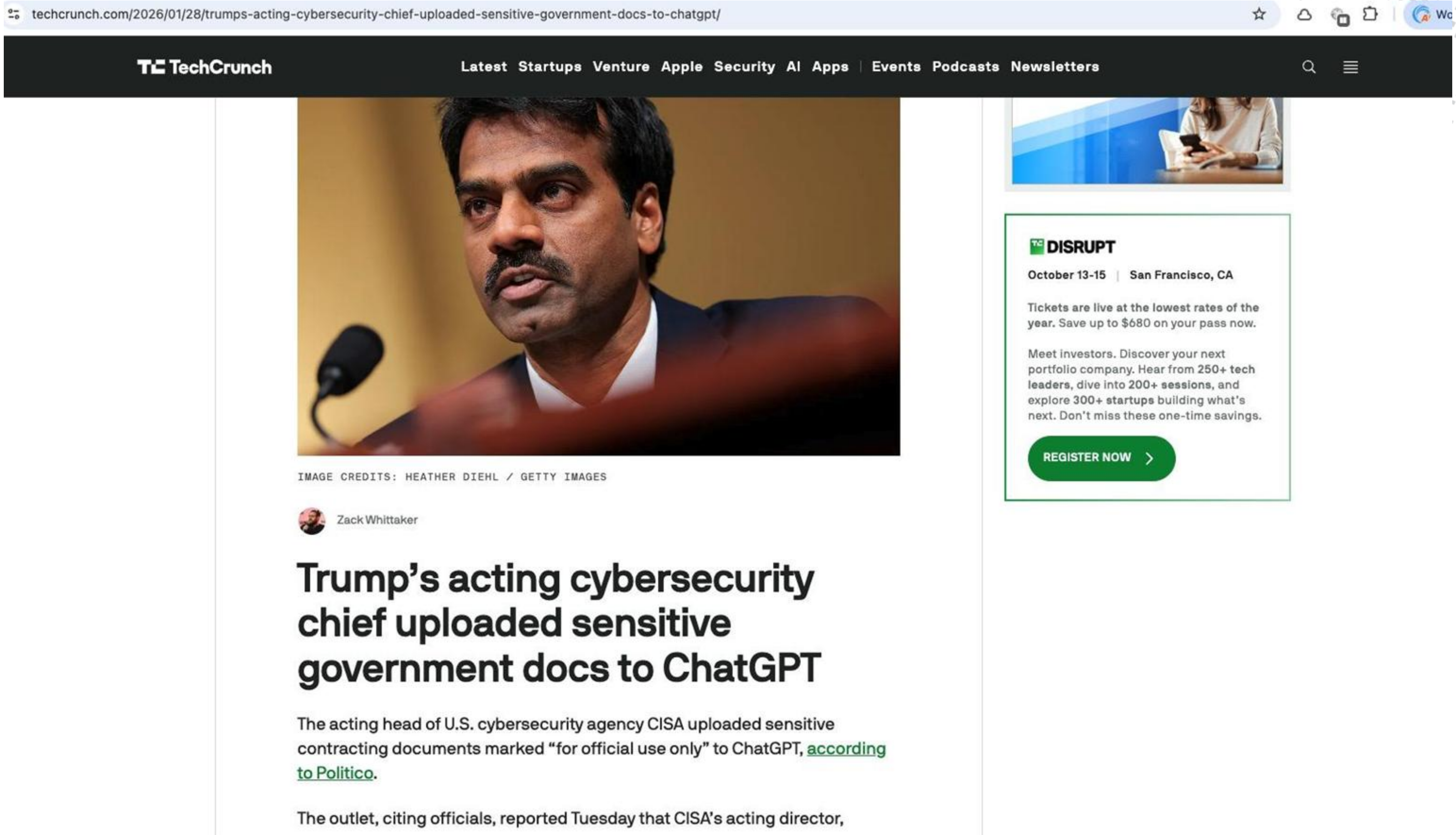

techcrunch.com/2026/01/28/trumps-acting-cybersecurity-chief-uploaded-sensitive-government-docs-to-chatgpt/
TechCrunch
Latest Startups Venture Apple Security AI Apps | Events Podcasts Newsletters

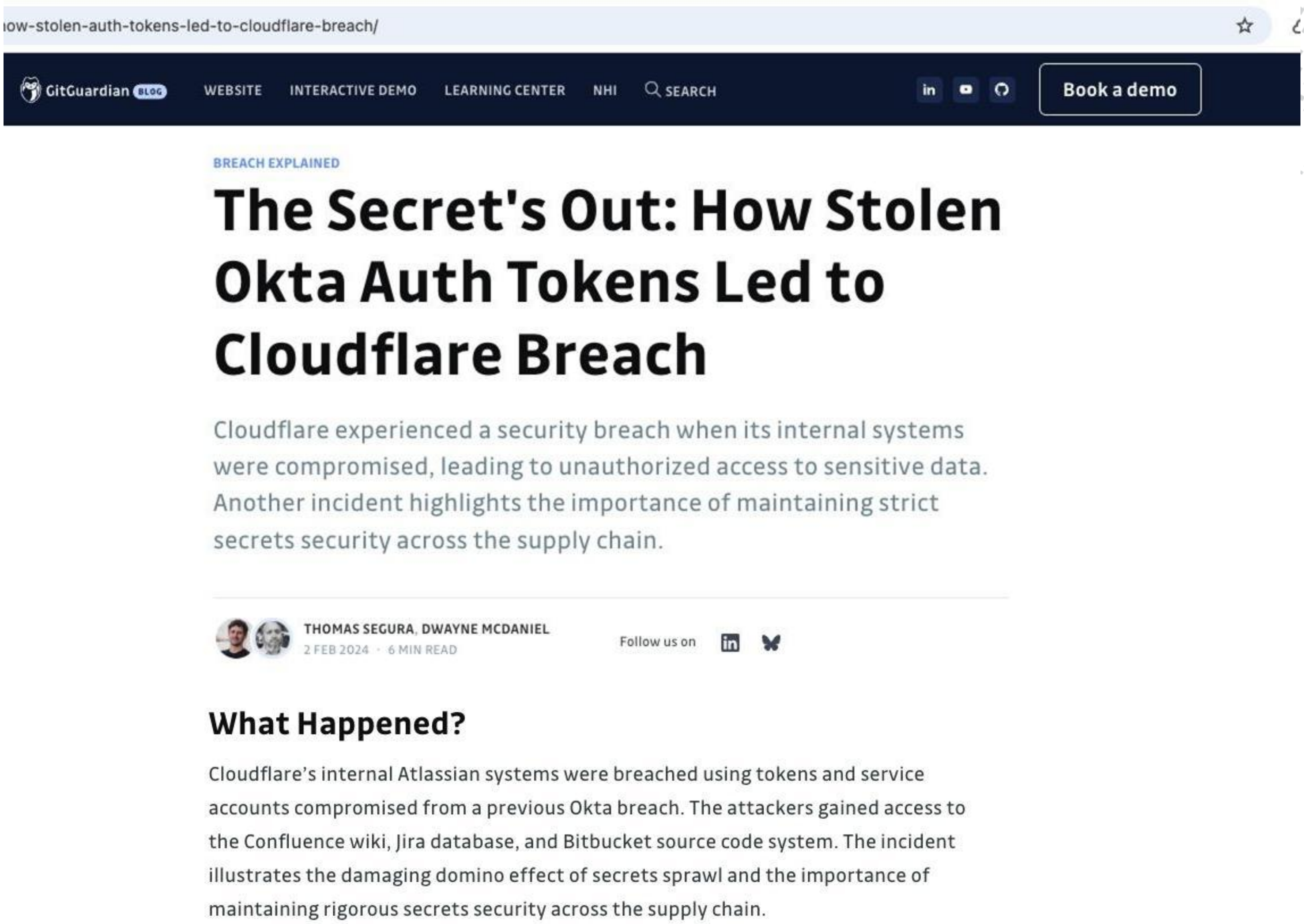
IMAGE CREDITS: HEATHER DIEHL / GETTY IMAGES

Zack Whittaker

Trump's acting cybersecurity chief uploaded sensitive government docs to ChatGPT

The acting head of U.S. cybersecurity agency CISA uploaded sensitive contracting documents marked "for official use only" to ChatGPT, according to Politico.

The outlet, citing officials, reported Tuesday that CISA's acting director,



BREACH EXPLAINED

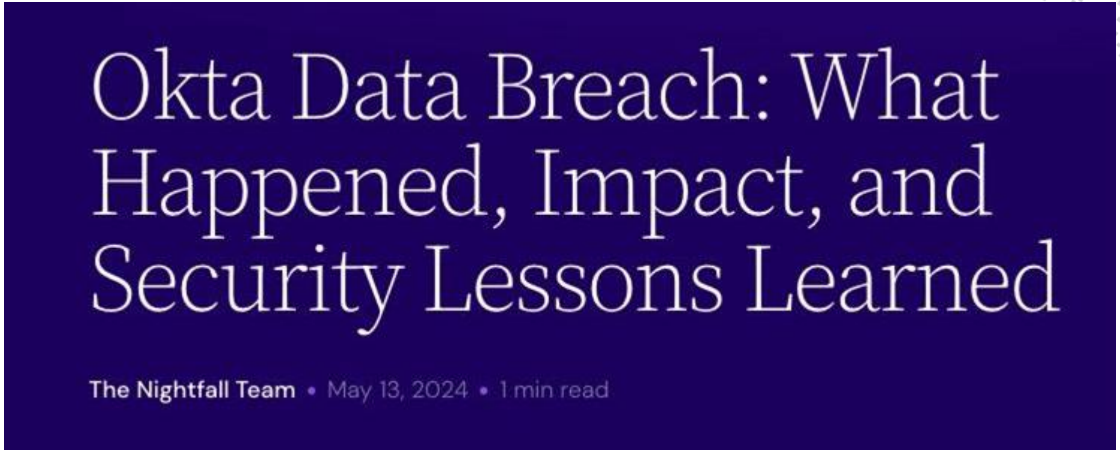
The Secret's Out: How Stolen Okta Auth Tokens Led to Cloudflare Breach

Cloudflare experienced a security breach when its internal systems were compromised, leading to unauthorized access to sensitive data. Another incident highlights the importance of maintaining strict secrets security across the supply chain.

THOMAS SEGURA, DWAYNE MCDANIEL
2 FEB 2024 · 6 MIN READ

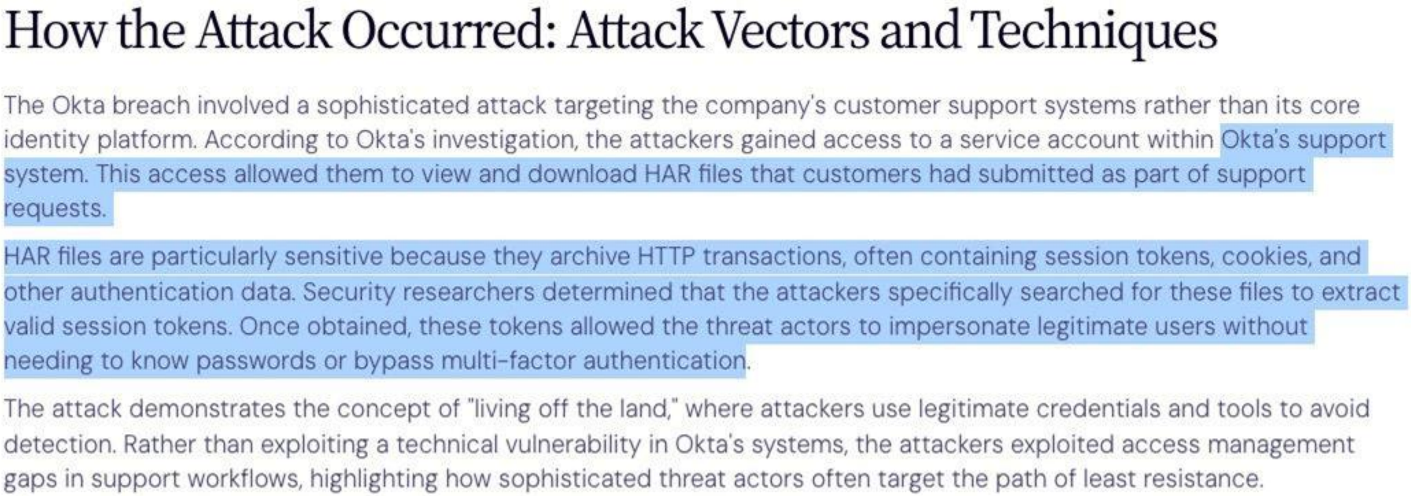
What Happened?

Cloudflare's internal Atlassian systems were breached using tokens and service accounts compromised from a previous Okta breach. The attackers gained access to the Confluence wiki, Jira database, and Bitbucket source code system. The incident illustrates the damaging domino effect of secrets sprawl and the importance of maintaining rigorous secrets security across the supply chain.



Okta Data Breach: What Happened, Impact, and Security Lessons Learned

The Nightfall Team • May 13, 2024 • 1 min read

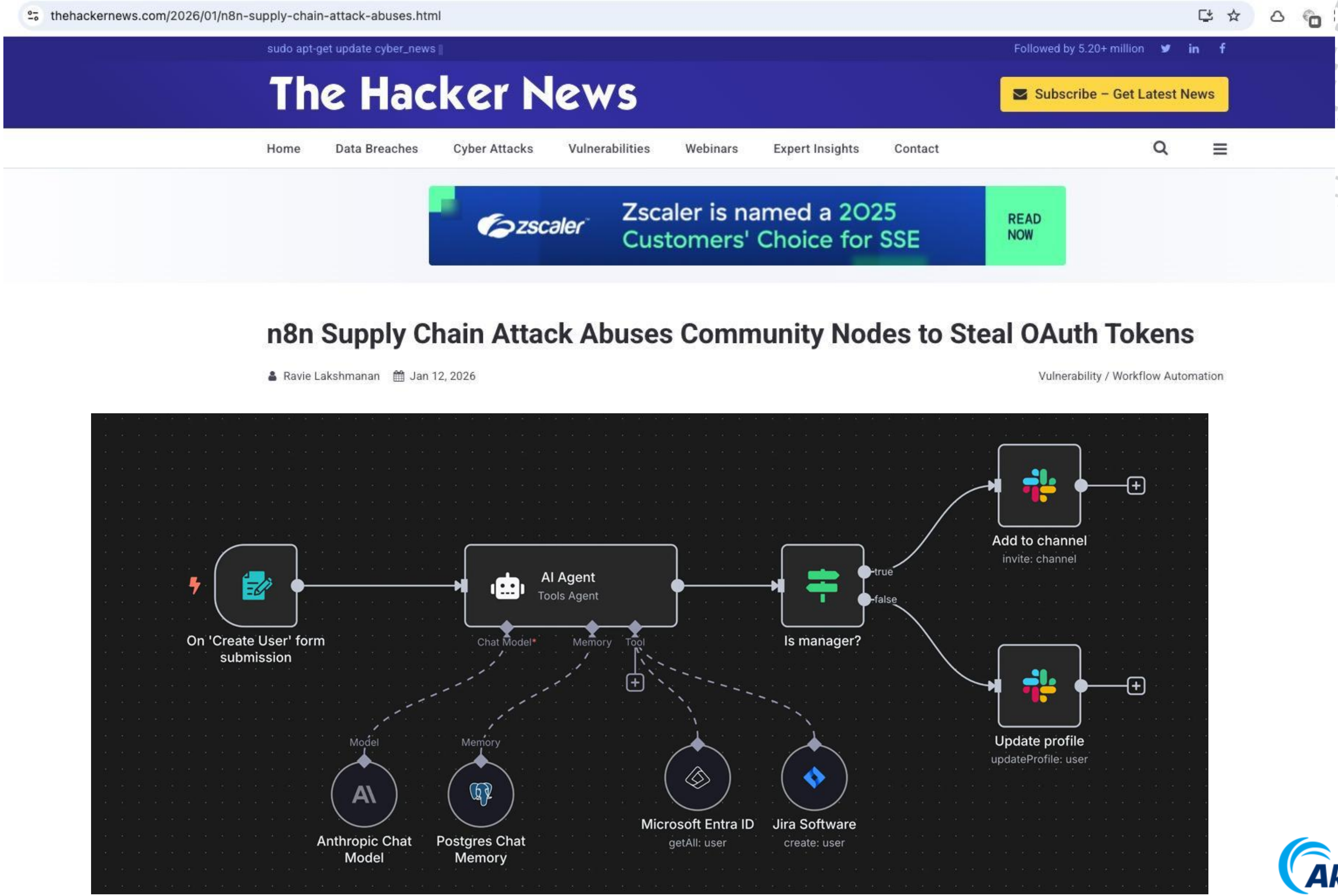


How the Attack Occurred: Attack Vectors and Techniques

The Okta breach involved a sophisticated attack targeting the company's customer support systems rather than its core identity platform. According to Okta's investigation, the attackers gained access to a service account within Okta's support system. This access allowed them to view and download HAR files that customers had submitted as part of support requests.

HAR files are particularly sensitive because they archive HTTP transactions, often containing session tokens, cookies, and other authentication data. Security researchers determined that the attackers specifically searched for these files to extract valid session tokens. Once obtained, these tokens allowed the threat actors to impersonate legitimate users without needing to know passwords or bypass multi-factor authentication.

The attack demonstrates the concept of "living off the land," where attackers use legitimate credentials and tools to avoid detection. Rather than exploiting a technical vulnerability in Okta's systems, the attackers exploited access management gaps in support workflows, highlighting how sophisticated threat actors often target the path of least resistance.



thehackernews.com/2026/01/n8n-supply-chain-attack-abuses.html

sudo apt-get update cyber_news

Followed by 5.20+ million

The Hacker News

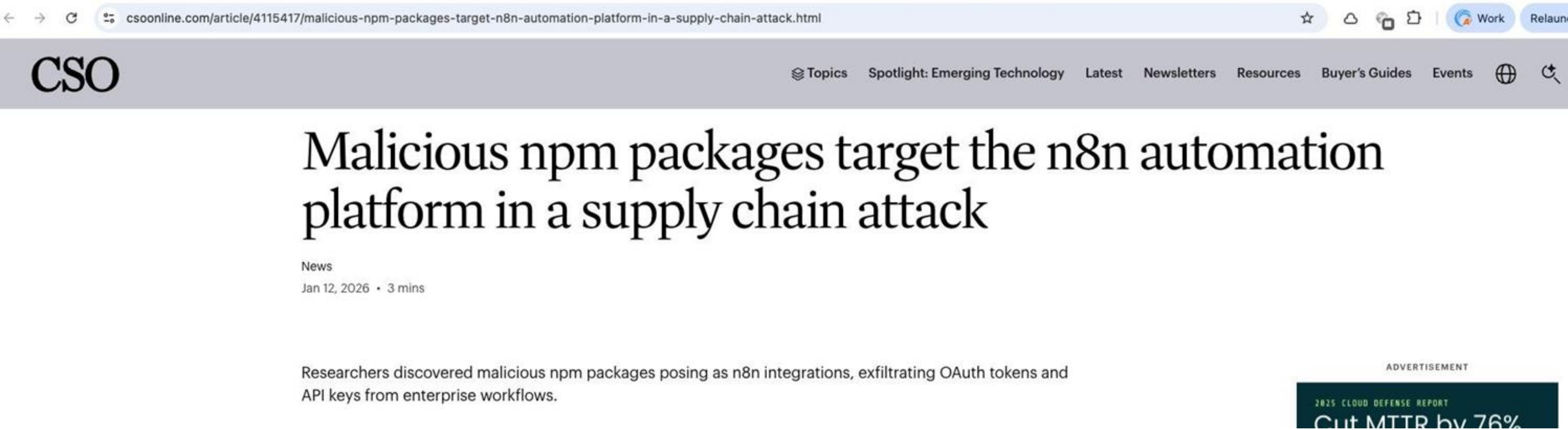
Subscribe – Get Latest News

Home | Data Breaches | Cyber Attacks | Vulnerabilities | Webinars | Expert Insights | Contact

n8n Supply Chain Attack Abuses Community Nodes to Steal OAuth Tokens

Ravie Lakshmanan Jan 12, 2026

Vulnerability / Workflow Automation



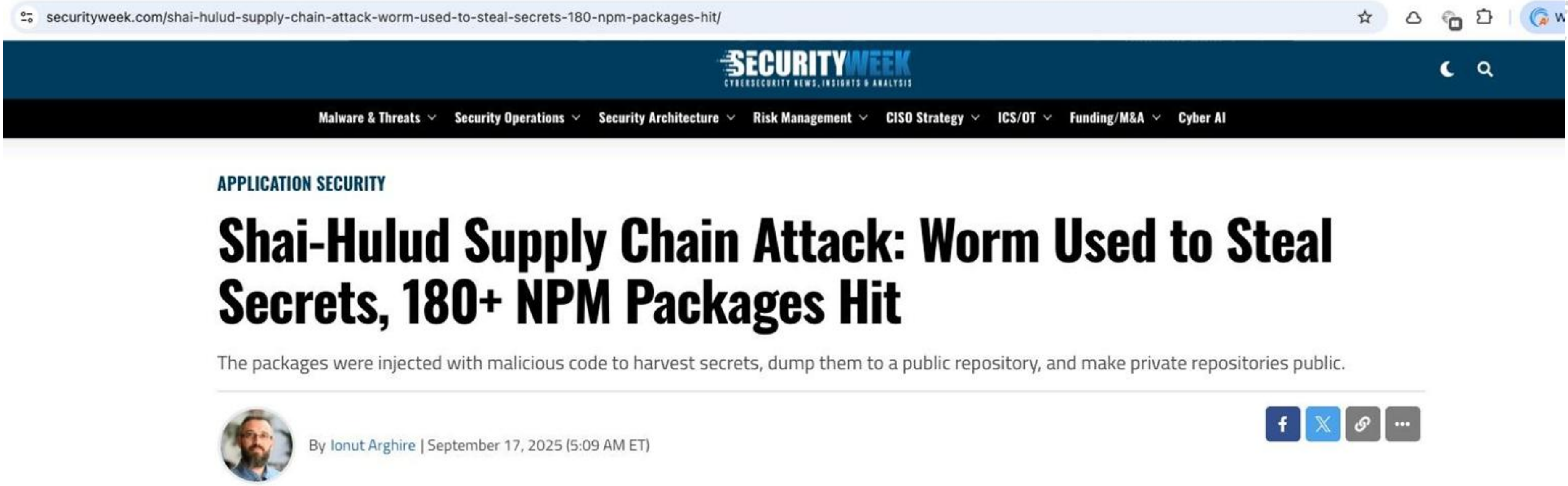
csoonline.com/article/4115417/malicious-npm-packages-target-n8n-automation-platform-in-a-supply-chain-attack.html
Work
Relaun
CSO
Topics
Spotlight: Emerging Technology
Latest
Newsletters
Resources
Buyer's Guides
Events

Malicious npm packages target the n8n automation platform in a supply chain attack

News

Jan 12, 2026 • 3 mins

Researchers discovered malicious npm packages posing as n8n integrations, exfiltrating OAuth tokens and API keys from enterprise workflows.

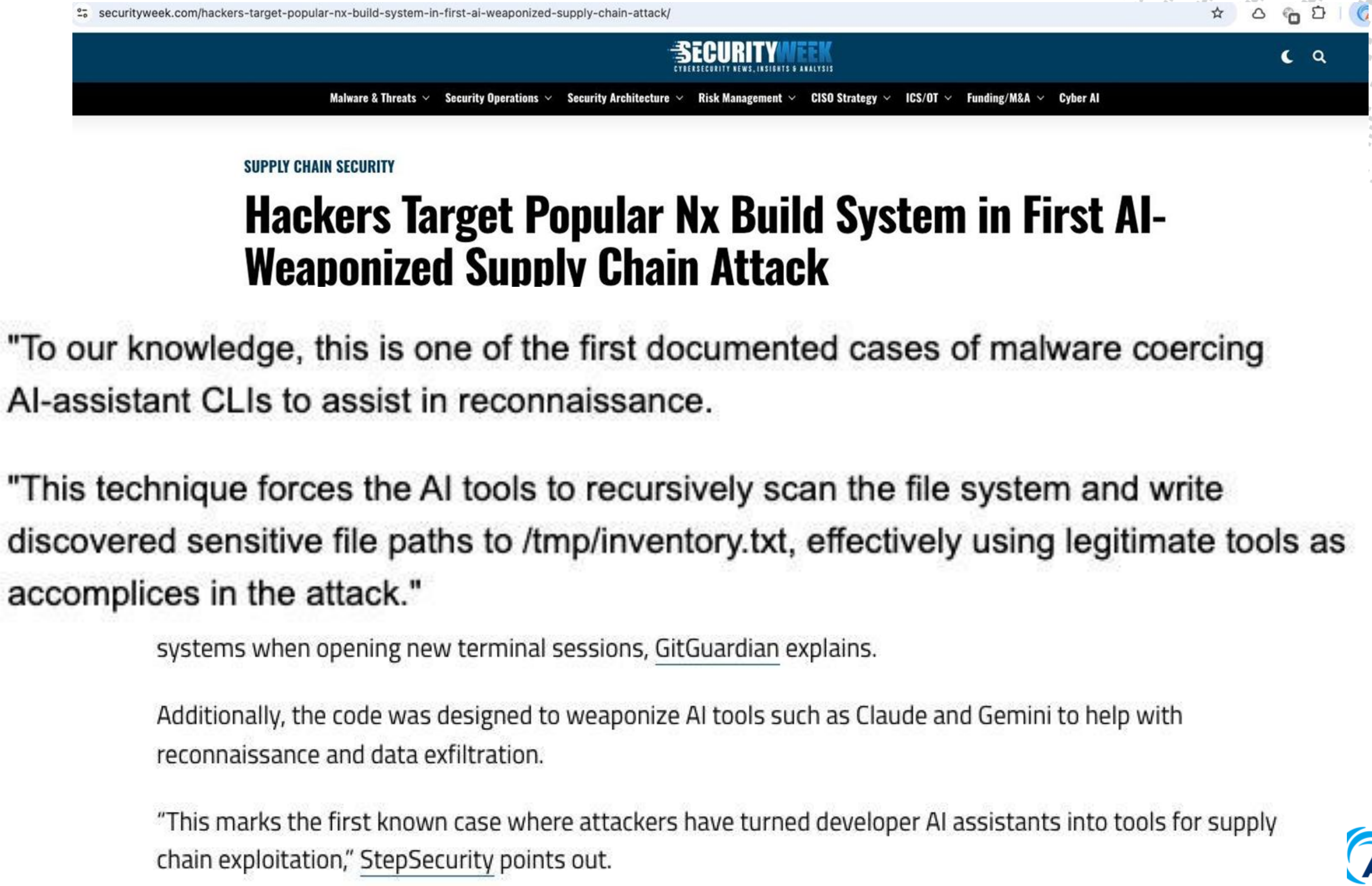

securityweek.com/shai-hulud-supply-chain-attack-worm-used-to-steal-secrets-180-npm-packages-hit/
SECURITYWEEK
CYBERSECURITY NEWS, INSIGHTS & ANALYSIS
Malware & Threats
Security Operations
Security Architecture
Risk Management
CISO Strategy
ICS/OT
Funding/M&A
Cyber AI

APPLICATION SECURITY

Shai-Hulud Supply Chain Attack: Worm Used to Steal Secrets, 180+ NPM Packages Hit

The packages were injected with malicious code to harvest secrets, dump them to a public repository, and make private repositories public.

By Ionut Arghire | September 17, 2025 (5:09 AM ET)


securityweek.com/hackers-target-popular-nx-build-system-in-first-ai-weaponized-supply-chain-attack/
SECURITYWEEK
CYBERSECURITY NEWS, INSIGHTS & ANALYSIS
Malware & Threats
Security Operations
Security Architecture
Risk Management
CISO Strategy
ICS/OT
Funding/M&A
Cyber AI

SUPPLY CHAIN SECURITY

Hackers Target Popular Nx Build System in First AI-Weaponized Supply Chain Attack

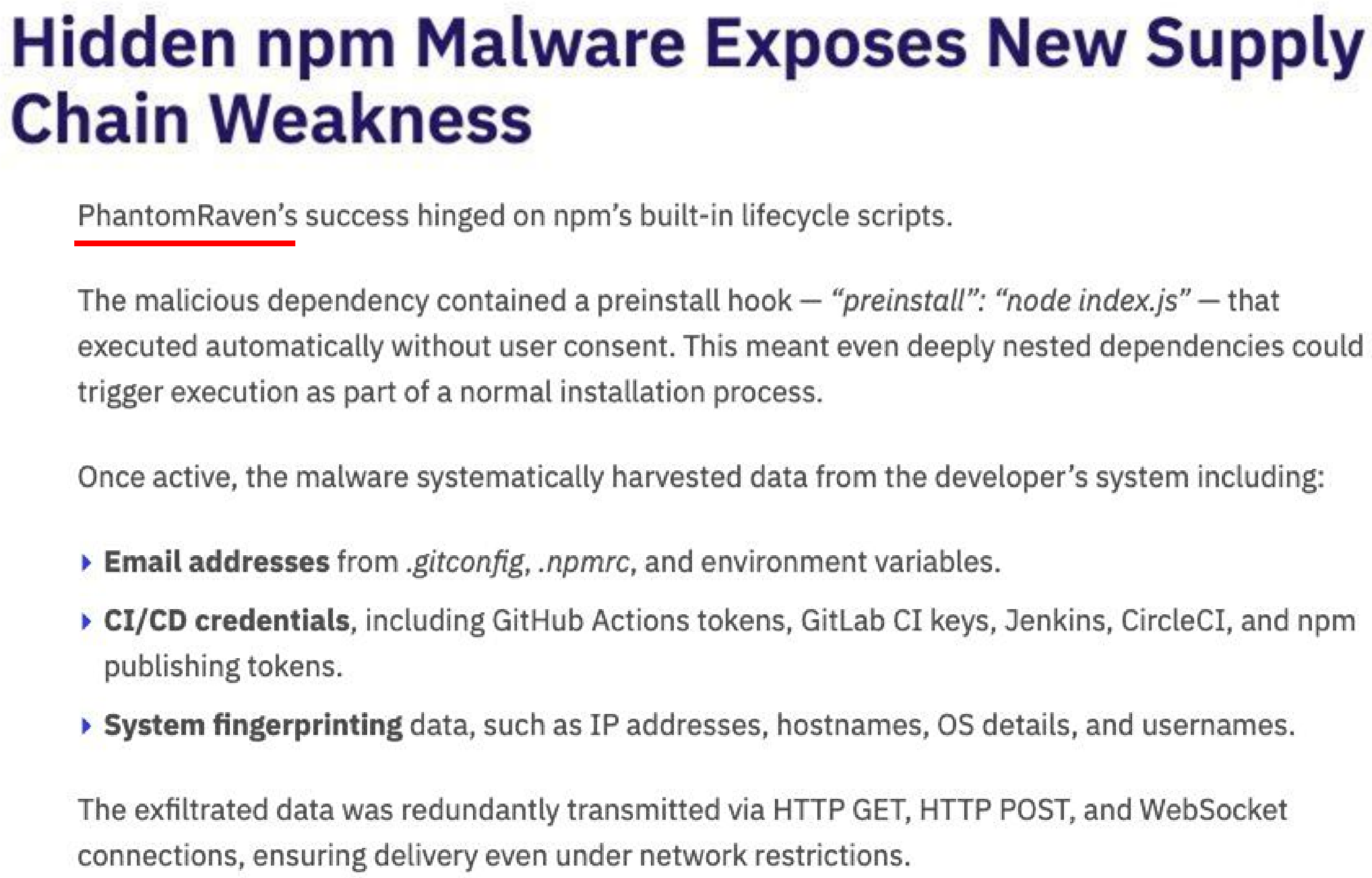
"To our knowledge, this is one of the first documented cases of malware coercing AI-assistant CLIs to assist in reconnaissance.

"This technique forces the AI tools to recursively scan the file system and write discovered sensitive file paths to /tmp/inventory.txt, effectively using legitimate tools as accomplices in the attack."

systems when opening new terminal sessions, GitGuardian explains.

Additionally, the code was designed to weaponize AI tools such as Claude and Gemini to help with reconnaissance and data exfiltration.

"This marks the first known case where attackers have turned developer AI assistants into tools for supply chain exploitation," StepSecurity points out.



Hidden npm Malware Exposes New Supply Chain Weakness

PhantomRaven's success hinged on npm's built-in lifecycle scripts.

The malicious dependency contained a preinstall hook — *"preinstall": "node index.js"* — that executed automatically without user consent. This meant even deeply nested dependencies could trigger execution as part of a normal installation process.

Once active, the malware systematically harvested data from the developer's system including:

- **Email addresses** from *.gitconfig*, *.npmrc*, and environment variables.
- **CI/CD credentials**, including GitHub Actions tokens, GitLab CI keys, Jenkins, CircleCI, and npm publishing tokens.
- **System fingerprinting** data, such as IP addresses, hostnames, OS details, and usernames.

The exfiltrated data was redundantly transmitted via HTTP GET, HTTP POST, and WebSocket connections, ensuring delivery even under network restrictions.

https://www.esecurityplanet.com/news/hidden-npm-malware-supply-chain/

Hidden npm Malware Exposes New Supply Chain Weakness

The Rise of Slopsquatting Attacks

Beyond the stealthy delivery mechanism, PhantomRaven introduced a novel social-engineering tactic called slopsquatting — a twist on traditional typosquatting.

Instead of mimicking existing package names, attackers registered plausible-sounding names that AI assistants like GitHub Copilot or ChatGPT might hallucinate on.

Examples include:

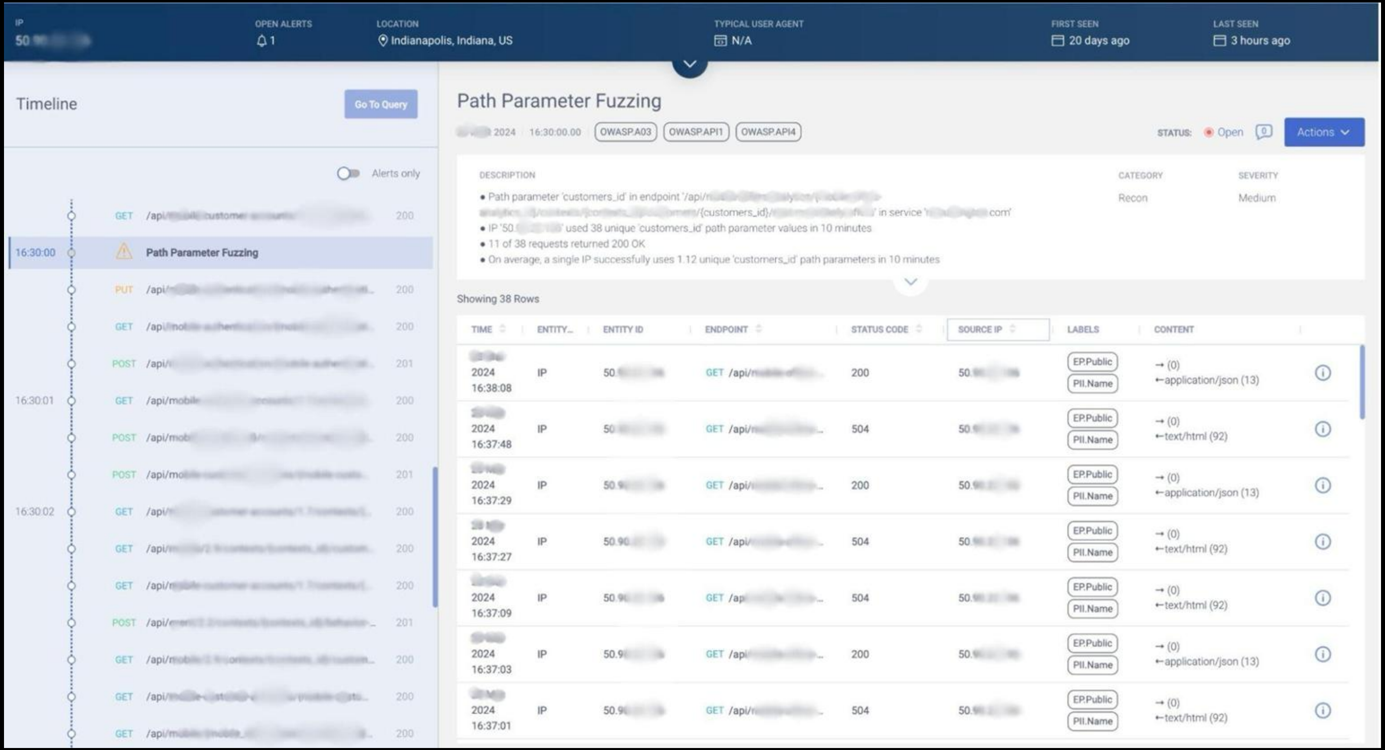
- *eslint-comments* instead of the legitimate *eslint-plugin-eslint-comments*
- *unused-imports* instead of *eslint-plugin-unused-imports*
- *Transform-react-remove-prop-types* instead of *babel-plugin-transform-react-remove-prop-types*

When AI suggested these nonexistent packages to developers, users unknowingly installed the malicious versions — demonstrating how artificial intelligence can unintentionally amplify supply-chain risks.

https://www.esecurityplanet.com/news/hidden-npm-malware-supply-chain/

Our approach:
Behavioural analysis

Example: Suspicious New Account Creation



IP 50 [redacted] | OPEN ALERTS 1 | LOCATION Indianapolis, Indiana, US | TYPICAL USER AGENT N/A | FIRST SEEN 20 days ago | LAST SEEN 3 hours ago

Timeline

```
         GET   /api/[redacted]customer[redacted]   200
16:30:00       Path Parameter Fuzzing
         PUT   /api/[redacted]                     200
         GET   /api/[redacted]                     200
         POST  /api/[redacted]                     201
16:30:01 GET   /api/mobile[redacted]               200
         POST  /api/mob[redacted]                  200
         POST  /api/mo[redacted]                   201
16:30:02 GET   /api/[redacted]                     200
         GET   /api/[redacted]                     200
         GET   /api/[redacted]                     200
         POST  /api/[redacted]                     201
         GET   /api/[redacted]                     200
         GET   /api/[redacted]                     200
         GET   /api/[redacted]                     200
```

Path Parameter Fuzzing

[redacted] 2024 | 16:30:00.00 | OWASP.A03 | OWASP.API1 | OWASP.API4

STATUS: Open

DESCRIPTION

- Path parameter 'customers_id' in endpoint '/api/[redacted]/{customers_id}/[redacted]' in service '[redacted]com'
- IP '50.[redacted]' used 38 unique 'customers_id' path parameter values in 10 minutes
- 11 of 38 requests returned 200 OK
- On average, a single IP successfully uses 1.12 unique 'customers_id' path parameters in 10 minutes

CATEGORY: Recon

SEVERITY: Medium

Showing 38 Rows

TIME	ENTITY...	ENTITY ID	ENDPOINT	STATUS CODE	SOURCE IP	LABELS	CONTENT
[redacted] 2024 16:38:08	IP	50.[redacted]	GET /api/[redacted]	200	50.[redacted]	EP.Public PII.Name	→ (0) ←application/json (13)
[redacted] 2024 16:37:48	IP	50[redacted]	GET /api/[redacted]	504	50.[redacted]	EP.Public PII.Name	→ (0) ←text/html (92)
[redacted] 2024 16:37:29	IP	50.9[redacted]	GET /api/[redacted]	200	50.[redacted]	EP.Public PII.Name	→ (0) ←application/json (13)
[redacted] 2024 16:37:27	IP	50.90[redacted]	GET /api/[redacted]	504	50.[redacted]	EP.Public PII.Name	→ (0) ←text/html (92)
[redacted] 2024 16:37:09	IP	50.9[redacted]	GET /ap[redacted]	504	50.[redacted]	EP.Public PII.Name	→ (0) ←text/html (92)
[redacted] 2024 16:37:03	IP	50.9[redacted]	GET /api[redacted]	200	50.[redacted]	EP.Public PII.Name	→ (0) ←application/json (13)
[redacted] 2024 16:37:01	IP	50.9[redacted]	GET /api/[redacted]	504	50.[redacted]	EP.Public PII.Name	→ (0) ←text/html (92)

1 Path Parameter Fuzzing alert

2 Abnormal fuzzing of {customer_id} path parameter

3 Correlated to suspicious new account creations

Key AI Security Challenges

Shadow AI

Discover your complete AI footprint - including rogue, legacy, shadow, zombie, etc.

Vulnerable AI tools

Prevent agentic vulnerabilities and misconfigurations

Agent Abuse

Stop business logic abuse such as data scraping or data exfiltration using behavioral analytics.

Tomorrow’s Focus

Today’s Focus

What this means for you

- Assume AI agents behave like insiders
- Treat automation as privileged
- Monitor behaviour, not just access
- Expect persistence, not attacks

Thank you