

EXECUTIVE PAPER · AI

Annual Security Tests Cannot Keep Pace With AI

Attackers now use AI to find and exploit weaknesses within days of disclosure, and sometimes before it. Boards and executives should fund security testing that runs every week and reports against the standards regulators already cite.

AUTHOR

Chris Watkins

SERVICE

AI

FOR

CEO, CTO, Board · 12 min read

AT A GLANCE

The problem. Google's Mandiant estimates that attackers in 2025 exploited vulnerabilities a week, on average, before a patch existed. In 2025 a state-backed group used an AI coding agent to carry out most of an espionage campaign against about 30 organizations. Meanwhile companies fixed only about a quarter of known exploited flaws, and took a median of 43 days to do it.

Why it persists. Security assurance is still bought as an annual event: an audit, a penetration test, a report filed for the auditors. Findings arrive months after the weakness appeared and are fixed in order of severity score instead of evidence of attack. AI applications and AI coding tools are usually outside the test scope entirely.

The recommendation. Fund a standing program that inventories what you expose, scans code and dependencies on every change, fixes anything on CISA's exploited list within days, tests web and AI applications against OWASP, and reports exposure to the board monthly. Map it to NIST CSF 2.0, the NIST AI Risk Management Framework, and ISO/IEC 27001 and 42001 so the same work satisfies auditors and regulators.

The Decision in Front of You

Most companies test their security once a year: an outside penetration test, an audit sample, and a remediation plan that runs into the next quarter. The decision is whether that cycle still matches the speed of the people attacking you.

Mandiant's M-Trends 2026 report, drawn from its 2025 investigations, put the mean time to exploit a vulnerability at an estimated negative seven days, meaning exploitation typically began before a fix was available [1]. Exploits were the most common way into a network for the sixth year running, at 32% of intrusions [1]. VulnCheck found that 28.96% of the vulnerabilities it confirmed as exploited in 2025 were attacked on or before the day they were publicly disclosed, up from 23.6% in 2024 [2].

Defenders are moving the other way. Coverage of Verizon's 2026 Data Breach Investigations Report says exploitation of vulnerabilities was the initial access route in 31% of breaches, that organizations fully fixed only 26% of the critical flaws on CISA's known-exploited list, and that the median time to fix rose from 32 to 43 days [3].

What's Happening

AI has shortened every step of an attack

In mid-September 2025 Anthropic detected a campaign it attributed with high confidence to a Chinese state-sponsored group, tracked as GTG-1002 [4]. The attackers pointed Anthropic's Claude Code tool at roughly 30 targets, including technology firms, banks, chemical manufacturers, and government agencies, and succeeded in a small number of cases [4]. The AI did an estimated 80% to 90% of the work, making thousands of requests, often several per second [4].

Malware now calls AI models mid-attack. Google's Threat Intelligence Group reported in November 2025 that PROMPTSTEAL, used by the Russian government-backed group APT28 against Ukraine, asks a language model to write the commands it runs on victim machines, the first such malware Google had seen in live operations [5].

Vulnerability discovery has changed most. In April 2026 Anthropic said an unreleased model, Claude Mythos Preview, had found thousands of high-severity vulnerabilities, including a 27-year-old flaw in OpenBSD, and warned that such capabilities would soon spread beyond defenders [6]. CrowdStrike's 2026 Threat Hunting Report says frontier AI is collapsing exploit windows and found China-nexus groups exploiting flaws 24 hours after a working proof of concept appeared [7].

The trend predates these tools. Mandiant put the average time to exploit at 63 days in 2018 and 2019 and 5 days in 2023 [9]. Rapid7 counted 146 exploited high- and critical-severity vulnerabilities in 2025, double the 71 of 2024 [8].

Four cases from 2024 and 2025

GTG-1002, 2025. The campaign above is the first documented case of an AI agent doing most of an intrusion's technical work [4]. No cost has been published. A few operators ran reconnaissance and exploitation against 30 organizations in parallel, with people stepping in at four to six decisions per campaign [4]. The agent sometimes invented credentials, which Anthropic called an obstacle to fully autonomous attacks [4]. The lesson is pace: one small team now works at the scale of many.

Arup, 2024. An employee in the Hong Kong office of the engineering firm Arup joined a video call with what appeared to be the company's chief financial officer and other colleagues [10]. All were deepfakes. He made 15 transfers totaling HK$200 million, about US$25.6 million, before checking with the UK head office [10]. The approval steps in a payment process are attack surface and need testing like any system.

Nx "s1ngularity," August 2025. An attacker exploited a flaw in an automated build workflow of Nx, a popular open-source build tool, stole its publishing token, and released malicious versions to the npm registry on August 26, 2025 [11]. Installing them ran a script that collected credentials and posted them to a public repository in the victim's own GitHub account [11]. The script also ran the developer's AI

coding tools, Claude Code, Gemini CLI, and Amazon Q, with flags that skip permission prompts, to hunt for secrets [12]. A build pipeline flaw is as exploitable as a product flaw, and AI tools widen what a bad package can reach.

Package hallucination, 2025. Researchers at the 2025 USENIX Security Symposium generated 576,000 code samples with 16 models and found that at least 5.2% of packages suggested by commercial models, and 21.7% by open-source models, did not exist [13]. An attacker who publishes malware under an invented name can wait for an AI assistant to install it, a technique called slopsquatting. Code scanning has to confirm every dependency is real and approved.

SIDEBAR: THE BASICS STILL DECIDE THE OUTCOME

Mandiant does not count 2025 as the year breaches came directly from AI; most intrusions still began with human and systemic failures [1]. AI helps attackers find an unpatched server or an over-permissioned token faster. An annual test finds the same weaknesses, months later.

Why the Usual Approach Fails

An annual penetration test is a snapshot. It tells you what a skilled team found during two or three weeks of one year. A flaw introduced the week after the test can stay invisible for eleven months, and the attackers described above need days.

Consider an illustration, the author's estimate. A company's internet-facing systems pick up one serious exploitable flaw a week, each fixed 30 days after discovery. If only the annual test finds them, the average flaw stays open about 212 days, or 11,050 days of exposure across 52 flaws. Quarterly scanning cuts the average to about 76 days. Weekly scanning with a seven-day fix target brings it to 10.5 days and the total to 546, about one-twentieth of the annual figure.

Several habits widen the gap.

Fixes are ranked by severity score, though many flaws rated critical are never attacked. CISA's Known Exploited Vulnerabilities catalog, usually called KEV, lists flaws confirmed as attacked in the wild, and CISA recommends that every organization use it as an input to patching priority [14]. It held 1,766 entries at the time of writing [14]. In June 2026 CISA replaced its earlier directive, BOD 22-01, with BOD 26-04, which sets federal agency deadlines by four factors: whether the system is exposed to the internet, whether the flaw is on the KEV list, whether exploitation can be automated, and how much control an attacker would gain [15]. The tightest tier requires a fix within three days plus a check for signs the system was already compromised [15]. The directive binds only federal civilian agencies, but its logic applies to any company.

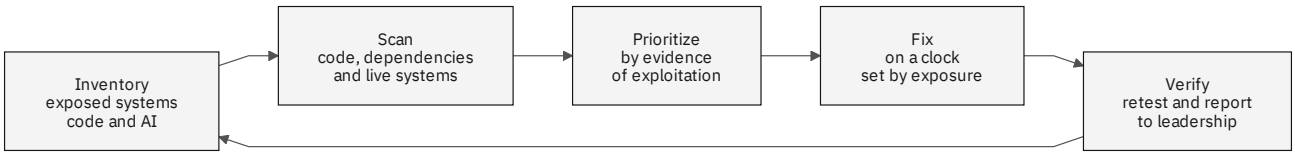
The inventory is incomplete. A test covers only listed systems, and a forgotten site or a project team's cloud service is often missing. BOD 26-04 requires agencies to tag every publicly exposed asset for this reason [15].

AI is outside the scope. Test plans rarely include the chatbot connected to the customer database, the AI coding tools on developer laptops, or the packages those tools install.

The Framework: The Exposure Loop

The Exposure Loop treats security testing as a control that runs continuously. Each stage has an owner and a clock, and the loop restarts every time code, infrastructure, or a vendor changes.

Exhibit 1. The Exposure Loop: every stage runs weekly, with an owner and a deadline



Source: author's framework, adapted from NIST CSF 2.0 [16] and CISA BOD 26-04 [15].

Inventory. Keep a live list of everything reachable from the internet, every code repository, every third-party package, and every AI system, model, and agent in use. NIST's Cybersecurity Framework 2.0 starts with this: asset inventories in its Identify function, and supplier risk under a new Govern function [16].

Scan. Code and dependency scanning run on every change, before code ships. Active scanning probes exposed systems weekly from the outside, as an attacker would. Web applications are tested against OWASP's Application Security Verification Standard, version 5.0.0 since May 2025 [21], using its Web Security Testing Guide [22]. AI applications are tested against the OWASP Top 10 for LLM Applications 2025, which starts with prompt injection and covers supply chain, poisoning, and excessive agency [23].

Prioritize. Rank findings by evidence of attack. A flaw on the KEV list, on an exposed system, that can be exploited automatically goes first [14][15]. Severity scores break ties.

Fix. Set fix deadlines by tier and publish them. A reasonable private-sector version of the federal model is three days for exploited flaws on exposed systems, two weeks for other exploited flaws, and the normal release cycle for the rest. Where a fix is not available, remove the system from the internet or put a compensating control in front of it.

Verify. Retest every fix. A few times a year, commission human penetration tests of the most valuable systems and AI applications, aimed at what automation misses, such as business logic and approval flows. Report open exposure to leadership monthly.

Each stage maps to a standard auditors recognize

Exhibit 2. The same work produces evidence for NIST, OWASP, ISO, and regulators

Stage	What to run	Standards it satisfies
Inventory	Asset, repository, dependency, and AI system inventory	NIST CSF 2.0 Identify and Govern [16]; ISO/IEC 42001 AI management system [26]
Scan: code	Static analysis, dependency and secret scanning, package allow lists	NIST SP 800-218A, the Secure Software Development Framework profile for generative AI [17]; CISA Secure by Design [25]
Scan: live systems	Weekly external active scanning; authenticated internal scanning	NIST CSF 2.0 ID.RA-01 vulnerability identification and DE.CM continuous monitoring [16]
Scan: applications	OWASP ASVS 5.0.0 requirements; WSTG test cases [21][22]	ISO/IEC 27001:2022 information security management [27]
Scan: AI systems	OWASP LLM Top 10 2025 tests; red teaming mapped to MITRE ATLAS [23] [24]	NIST AI RMF and Generative AI Profile [18][19]; NIST AI 100-2 attack taxonomy [20]
Prioritize and fix	KEV-driven deadlines by exposure tier	CISA KEV and BOD 26-04 model [14][15]
Verify and report	Retest, targeted penetration tests, monthly board report	SEC Item 106 risk management and board oversight disclosure [28]

Source: author's mapping of the cited standards.

For AI, NIST's AI Risk Management Framework (AI 100-1) sets four functions, Govern, Map, Measure, and Manage [18], and its 2024 Generative AI Profile (AI 600-1) adds risks specific to generative systems [19]. NIST AI 100-2 E2025 gives testers and auditors shared names for attacks such as poisoning and evasion [20], and MITRE ATLAS catalogs real attacker tactics against AI systems [24].

Regulators are asking for the same evidence

Public companies in the United States must disclose a material cybersecurity incident on Form 8-K within four business days of deciding it is material, and must describe in the annual 10-K how they assess and manage cyber risk and how the board oversees it [28]. A monthly exposure report gives the board accurate material for that filing.

For companies building or deploying AI in Europe, under the EU AI Act as amended on July 27, 2026, transparency obligations apply from August 2, 2026, and obligations for standalone high-risk systems move to December 2, 2027, with AI embedded in regulated products following on August 2, 2028 [29].

ISO/IEC 42001, published in 2023, gives companies a certifiable management system for AI that sits beside ISO/IEC 27001 [26][27].

Costs, Risks, and Tradeoffs

A continuous program costs more than an annual test: licenses, weekly triage, and engineering time spent on fixes. The exposure model above shows what that buys.

Exhibit 3. Tradeoffs of continuous testing

Area	Cost or risk	Mitigation
Tooling	Scanner, dependency, and secret-scanning licenses grow with the number of repositories and assets	Start with internet-facing systems and the repositories that ship to them
Alert fatigue	Scanners report thousands of findings, most never exploited	Gate on KEV status and exposure; suppress findings with a documented reason
Engineering time	Fix deadlines pull developers off planned work	Reserve a fixed share of each sprint; automate dependency updates
False confidence	Clean scans miss logic flaws, social engineering, and payment fraud like Arup [10]	Keep targeted human testing and test approval processes for payments
AI tools in development	Coding assistants install packages and run commands with developer permissions [12] [13]	Allow lists for packages; no permission-skipping modes on machines with production credentials
When this is wrong	A small company running almost entirely on SaaS has little code or infrastructure of its own to scan	Require vendors' security attestations, turn on MFA and logging, and test the payment approval process

Source: author's analysis.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	Inventory exposed systems, repositories, AI systems, and AI coding tools. Fix or isolate every KEV match. Turn off permission-skipping modes.	CISO with CTO	Zero KEV-listed flaws on internet-facing systems; no AI tool runs unattended with production credentials
First 90 days	Add code, dependency, and secret scanning to every build. Start weekly external active scanning. Publish fix deadlines by tier.	Application security lead	Every production repository scanned on every change; first weekly scan report
First 90 days	Test each customer-facing AI application against the OWASP LLM Top 10. Confirm payment approvals need a second channel.	Security lead with finance	Test report per AI application; out-of-band check on payments above a set amount
First 365 days	Report open KEV flaws, median days to fix, and inventory coverage monthly. Commission targeted penetration tests. Map evidence to NIST and ISO.	CISO with risk or compliance	Exploited flaws on exposed systems fixed within the published deadline 95% of the time; audit evidence on file; quarterly board report

Source: author's recommendation.

Questions to Ask Your Team

1. How many systems do we expose to the internet, and how do we know the list is complete?
2. How many flaws on CISA's exploited list are open on those systems today, and for how long?
3. What is our median time to fix an exploited flaw, and what deadline have we published?
4. Which AI applications and agents do we run, and when were they last tested for prompt injection?
5. Which AI coding tools do developers use, and what credentials can those tools reach?
6. Could a convincing video call from our CFO move money without a second check?
7. What would we write in our 10-K about how we manage cyber risk, and what evidence supports it?

Closing

Attackers now exploit weaknesses in days, sometimes before a fix exists, and AI is making them faster. A program that looks once a year leaves those weaknesses open for months. Fund continuous testing that scans every change, fixes exploited flaws on a published clock, tests AI systems like any other

application, and reports open exposure to the board monthly. Mapped to NIST, OWASP, and ISO, the same work gives auditors and regulators the evidence they expect.

Sources

[1] Kutscher, J. (Mandiant, Google Cloud). "M-Trends 2026: Data, Insights, and Strategies From the Frontlines." Google Cloud Blog. 2026. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026/.

[2] VulnCheck. "State of Exploitation 2026." 2026. https://www.vulncheck.com/blog/state-of-exploitation-2026.

[3] eSecurity Planet. "2026 Verizon DBIR: The New Era of Cyber Threats." Coverage of the Verizon 2026 Data Breach Investigations Report. 2026. https://www.esecurityplanet.com/threats/2026-verizon-dbir-the-new-era-of-cyber-threats/.

[4] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." 2025. https://www.anthropic.com/news/disrupting-AI-espionage.

[5] Google Threat Intelligence Group. "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools." Google Cloud Blog. 2025. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools.

[6] Anthropic. "Project Glasswing." 2026. https://www.anthropic.com/glasswing.

[7] CrowdStrike. "2026 Threat Hunting Report." 2026. https://www.crowdstrike.com/en-us/resources/reports/threat-hunting-report/.

[8] Rapid7. "Rapid7 2026 Global Threat Landscape Report Shows Exploited High and Critical Severity Vulnerabilities Surged 105% as Attack Timelines Collapsed." Press release. 2026. https://www.rapid7.com/about/press-releases/rapid7-2026-global-threat-landscape-report-shows-exploited-high-and-critical-severity-vulnerabilities-surged-105-as-attack-timelines-collapsed.

[9] Mandiant, Google Cloud. "How Low Can You Go? An Analysis of 2023 Time-to-Exploit Trends." Google Cloud Blog. 2024. https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023.

[10] Fortune. "A deepfake 'CFO' tricked the British design firm behind the Sydney Opera House in $25 million scam." 2024. https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo.

[11] Nx (nrwl). "Malicious versions of Nx and some supporting plugins were published." GitHub Security Advisory GHSA-cxm3-wv7p-598c. 2025. https://github.com/nrwl/nx/security/advisories/GHSA-cxm3-wv7p-598c.

[12] Snyk. "Weaponizing AI Coding Agents for Malware in the Nx Malicious Package Security Incident." 2025. https://snyk.io/blog/weaponizing-ai-coding-agents-for-malware-in-the-nx-malicious-package.

[13] Spracklen, J., Wijewickrama, R., Sakib, A. H. M. N., Maiti, A., Viswanath, B., and Jadliwala, M. "We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs." USENIX Security 2025. arXiv:2406.10279. 2025. https://arxiv.org/abs/2406.10279.

[14] Cybersecurity and Infrastructure Security Agency. "Known Exploited Vulnerabilities Catalog." https://www.cisa.gov/known-exploited-vulnerabilities-catalog.

[15] Cybersecurity and Infrastructure Security Agency. "BOD 26-04: Prioritizing Security Updates Based on Risk." Binding Operational Directive 26-04. 2026. https://www.cisa.gov/news-events/directives/bod-26-04-prioritizing-security-updates-based-risk.

[16] National Institute of Standards and Technology. "The NIST Cybersecurity Framework (CSF) 2.0." NIST CSWP 29. 2024. https://doi.org/10.6028/NIST.CSWP.29.

[17] National Institute of Standards and Technology. "Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile." NIST SP 800-218A. 2024. https://csrc.nist.gov/pubs/sp/800/218/a/final.

[18] National Institute of Standards and Technology. "AI Risk Management Framework." NIST AI 100-1. 2023. https://www.nist.gov/itl/ai-risk-management-framework.

[19] National Institute of Standards and Technology. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." NIST AI 600-1. 2024. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence.

[20] National Institute of Standards and Technology. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations." NIST AI 100-2 E2025. 2025. https://csrc.nist.gov/pubs/ai/100/2/e2025/final.

[21] OWASP Foundation. "Application Security Verification Standard." Version 5.0.0. 2025. https://asvs.owasp.org/.

[22] OWASP Foundation. "Web Security Testing Guide." Version 4.2. https://owasp.org/www-project-web-security-testing-guide/.

[23] OWASP GenAI Security Project. "OWASP Top 10 for LLM Applications 2025." 2025. https://genai.owasp.org/llm-top-10/.

[24] MITRE. "ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." https://atlas.mitre.org/. Description and scope as summarized by Akamai, "MITRE ATLAS." 2026. https://www.akamai.com/blog/ai/mitre-atlas.

[25] Cybersecurity and Infrastructure Security Agency. "Secure by Design." https://www.cisa.gov/securebydesign.

[26] International Organization for Standardization. "ISO/IEC 42001:2023 Information technology: Artificial intelligence: Management system." 2023. https://www.iso.org/standard/42001.

[27] International Electrotechnical Commission Webstore. "ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection: Information security management systems: Requirements." 2022. https://webstore.iec.ch/publication/79694.

[28] U.S. Securities and Exchange Commission. "SEC Adopts Rules on Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure by Public Companies." Press Release 2023-139. 2023. https://www.sec.gov/newsroom/press-releases/2023-139.

[29] Legal 500. "EU Digital Omnibus on AI Enters Into Force." Regulation (EU) 2026/1744. 2026. https://www.legal500.com/intelligence/european-union/technology/eu-digital-omnibus-on-ai-enters-into-force.

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