AI Risk • Governance • Regulatory Intelligence

*Issue 001 | April 2026*

## This Issue

**Agentic offense is operational.** Anthropic disclosed the first vendor attributed nation state agentic intrusion campaign. DARPA proved offensive AI research can run in public with full oversight. Same capability class. Two governance outcomes. This issue unpacks what that means for your program.

uniticybermedia.com | inquiry@uniticybermedia.com

OPENING

## The vendor noticed before the victims did.

*In November 2025, Anthropic published a threat disclosure about a Chinese state linked actor that had automated most of an active intrusion campaign using Claude Code. Roughly 30 organizations across government, finance, and technology were in scope.*

*Here's the part to sit with. The targeted organizations didn't catch it. Their SOCs didn't catch it. Their EDR didn't catch it. The vendor's abuse monitoring caught it, terminated access, and called the victims. That is not a detection program. That is luck wearing a lanyard.*

*The same month, DARPA closed its AI Cyber Challenge and handed out $8.5 million to the top three teams for building autonomous systems that find and patch real vulnerabilities at machine speed. Same capability class. Two governance outcomes. That's what this issue is about.*

FEATURE

## Your detection clock is set to the wrong tempo.

*Think about how a building alarm works. It's calibrated to a burglar. A human pace, a human rhythm of doors and footsteps and pauses. Now picture the intruder as a small drone that sweeps every room before a person could climb one flight of stairs. The alarm is still working. It's just watching for a threat that is no longer the threat.*

*That's the governance failure underneath the GTG-1002 disclosure. Anthropic assessed that between 80 and 90 percent of the campaign's operational tasks were automated. The model handled reconnaissance, generated credential harvesting workflows, adapted exploits to specific targets, and produced reasoning for lateral movement. Humans set objectives and approved key decisions. The agent did most of the work.*

*Does your SOC run at that tempo? Most don't. CrowdStrike's 2026 Global Threat Report puts the average breakout time for AI assisted intrusions at 29 minutes. The industry average enterprise patch cycle is still 60 days, against a weaponization window that averaged 4.5 days across 2025 and 2026 incidents. Your detection and your response are both behind the curve by roughly an order of magnitude, and your board probably hasn't seen the math.*

*Here's where it gets interesting. The same capability class that's running offense against you is running defense in public, with the results on the record. DARPA's AI Cyber Challenge finished at DEF CON 33 in August 2025. Team Atlanta won $4 million. Trail of Bits took $3 million. Theori came in third at $1.5 million. Across the program, autonomous systems discovered 18 previously unknown vulnerabilities in production open source code and generated working patches for most of them inside the competition window.*

*The contrast isn't about which lab built a better agent. It's about what gets built around the agent. DARPA defined the scope before any capability deployed. Teams were vetted and registered. Findings went to upstream maintainers through coordinated disclosure. Winning systems were published. No surprise zero-day drops. A full record. That is NIST AI RMF MAP 1.1, GOVERN 1.1, and GOVERN 1.2 made operational, not cited in a policy binder.*

*The GTG-1002 targets had none of that. Their detection was calibrated to human pace adversaries. Their response runbooks serialized steps the agent parallelized. Their AI tool access controls didn't exist as a detection domain because the threat model underneath them predated agentic offense. NIST AI RMF GOVERN 1.4 requires organizations to maintain current awareness of the threat environment their systems operate in. A 2023 threat model applied to a 2025 adversary fails that requirement on its face.*

*The governance implication isn't that your organization needs to build offensive AI. It's that your vulnerability management program, your patch cadence, your third party pentest contracts, and your incident response assumptions were all designed for a threat timeline that no longer exists.*

*Treating detection as a static control is no longer a best practice. It's an audit finding waiting to land.*

RISK VISIBLE

## GTG-1002: When the vendor finds your breach before you do.

*On November 14, 2025, Anthropic disclosed that a Chinese state linked actor, tracked as GTG-1002, used Claude Code to automate an estimated 80 to 90 percent of an intrusion campaign against approximately 30 entities across government, finance, and technology. Anthropic's internal abuse monitoring caught the pattern. Access was terminated.*

*Affected parties were notified where possible. The victim organizations did not detect it themselves.*

*Stay with that for a second. Approximately 30 organizations, and the tell came from a vendor's telemetry, not from any of their security programs.*

*The specific control gaps map to framework requirements. NIST CSF 2.0 DE.CM (continuous monitoring) and DE.AE (adverse event analysis) define the detection function. Both assume a baseline of normal system behavior that an AI assisted adversary bypasses by never behaving normally for long enough to establish one. NIST IR 8596, the draft Cyber AI Profile released December 16, 2025, closes that gap directly by introducing control expectations for detecting AI assisted offensive activity against an organization. Regulated industry examiners are expected to begin citing it within two audit cycles.*

*The lesson is uncomfortable. If your detection program depends on a vendor noticing your breach and calling you, you don't have a detection program. You have a vendor relationship that happened to save you this time.*

*What to do with this. Pull your current detection architecture and ask whether it assumes human pace adversaries. Identity based anomaly detection needs a mean time to detect under 30 minutes, because that's the documented breakout tempo. Incident response playbooks written for human attackers serialize steps an agentic attacker runs in parallel. AI tool access controls are now a security perimeter, not a procurement concern. Who can access your AI platforms, under what conditions, with what monitoring, and who would notice if that monitoring went dark for a week?*

*The governance question is not whether AI will be used against you. It's whether your detection, response, and access controls are designed for that reality, or whether your next breach notification is going to arrive the same way GTG-1002's did.*

*This is our read through a GRC lens. The 80 to 90 percent automation figure is vendor attributed and has not been independently replicated by third party threat intelligence as of April 2026. Attribution of the campaign to a specific Chinese state program has not been confirmed in public reporting by any named national intelligence service we reviewed. Not legal advice, not regulatory gospel. Stress test this against your own detection program and decide what it means for your organization.*

ON THE RADAR

## What to Watch This Month

*Three items worth watching this month.*

- **NIST IR 8596, Draft Cyber AI Profile, Public Comment Open.** NIST released the draft profile on December 16, 2025. It's the first US government document that codifies defender expectations for monitoring AI assisted offensive activity against an organization, cross referenced to CSF 2.0 and AI RMF functions. Regulated industry examiners are expected to begin citing it within two audit cycles.
- **Insurance Market, AI Losses Now a Separate Pricing Category.** In January 2026, ISO updated its general liability language to exclude AI related losses unless separately endorsed. Coalition responded with an affirmative AI endorsement tied to an underwriting questionnaire covering AI enabled adversary detection readiness. Pull your policy. Read the exclusion language.
- **August 2, 2026. EU AI Act High Risk Enforcement.** High risk system requirements become enforceable for most organizations, including non-EU companies whose AI systems affect EU residents. Penalties reach €15 million or 3% of global annual turnover. If you haven't mapped your systems to the four tier risk architecture, the clock's running.

THE CLOSE

*Same capability class. Two governance outcomes.*

*DARPA ran offensive AI research in public with attested access, defined scope, coordinated disclosure, and the winning systems open sourced for defenders to use. A full record, with names attached.*

*GTG-1002 ran against organizations whose detection was tuned to the adversary their SOC was designed to catch a decade ago. The vendor noticed. The victims didn't.*

*The difference isn't capability. It's what got built around it, and when. Organizations that treat agentic offense as an IT problem will find out why that was the wrong call when the incident happens or the regulator asks. Organizations that treat it as a governance problem, with scope and access and accountability on paper before the capability shows up, will answer a different set of questions.*

*Which one is yours?*

*Reply and tell us what governance challenge you're navigating. We read every response, and it shapes what we cover next.*

***We do not rise to the level of our AI capabilities.***

***We fall to the level of our governance.***

*Build it in from the start. Everything else is just damage control.*

**Built in, not bolted on.**

SOURCES & NOTES

# Citations and verification notes

**Anthropic.** Public threat disclosure of GTG-1002 agentic intrusion campaign. November 14, 2025. The 80 to 90 percent automation figure is Anthropic's internal assessment and has not been independently replicated by third party threat intelligence as of April 2026.

**CrowdStrike.** 2026 Global Threat Report. Reference for 29-minute average breakout time on AI assisted intrusions.

**DARPA.** AI Cyber Challenge (AIxCC) final report and program materials, 2023 through 2025. Reference for program structure, prize distribution (Team Atlanta $4M, Trail of Bits $3M, Theori $1.5M), and coordinated disclosure of 18 previously unknown vulnerabilities in production open source code.

**NIST.** Cybersecurity Framework 2.0 (NIST CSWP 29). 2024. Reference for DE.CM and DE.AE categories of the Detect function.

**NIST.** Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. January 2023. Reference for GOVERN 1.1, GOVERN 1.2, GOVERN 1.4, and MAP 1.1.

**NIST.** Cyber AI Profile (NIST IR 8596, draft). December 16, 2025. First US government document codifying defender expectations for monitoring AI assisted offensive activity.

**SANS, Cloud Security Alliance, and OWASP GenAI.** Joint emergency strategy briefing on AI assisted exploit timeline compression. April 14, 2026. Reference for 4.5-day weaponization window and 60-day enterprise patch cycle.

**Insurance Services Office (ISO).** General liability exclusion language for AI related losses. January 2026. Coalition affirmative AI endorsement announcement, 2026.

**European Union.** Regulation (EU) 2024/1689 (EU AI Act), high risk system requirements. Enforceable from August 2, 2026.

*Attribution note: The GTG-1002 campaign's attribution to a specific Chinese state program has not been confirmed in public reporting by any named national intelligence service we reviewed for this issue.*