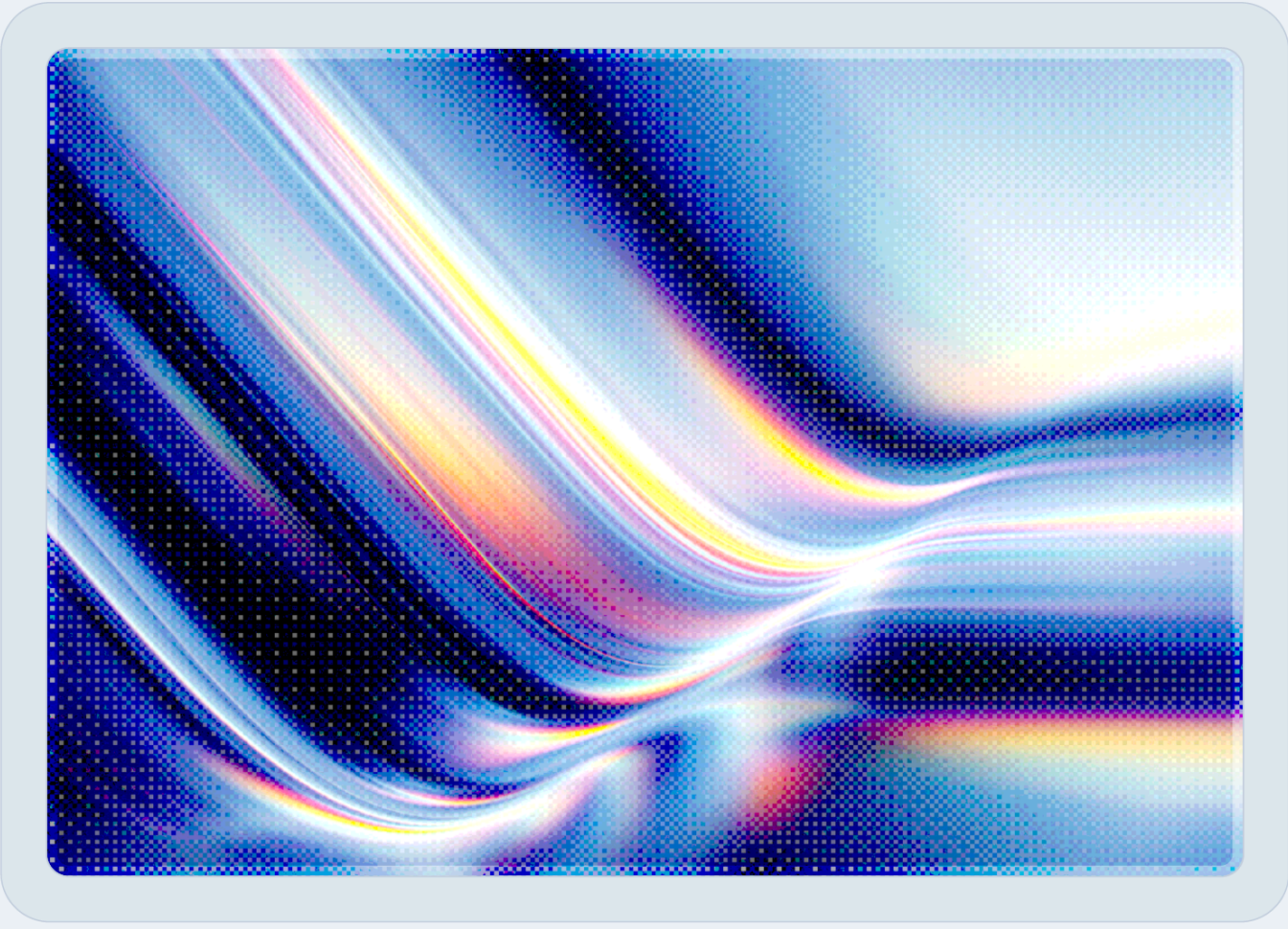
CSAI Foundation | Cloud Security Alliance

State-Level Tempo for Everyone: Commodity Agent Frameworks and Adversary Tiering

How open offensive agent harnesses are narrowing the operational gap between nation-state operators and everyone else

2026-11-13

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

Exactly one year after Anthropic disclosed GTG-1002, which it described as the first reported AI-orchestrated cyber espionage campaign [1], the most consequential development in the threat landscape, in CSA's assessment, is no longer that state actors use AI agents. It is that the same agent frameworks are now shared by intelligence services, criminal crews, and lone operators, and the gap in what each can achieve has narrowed. State-aligned actors have moved from proprietary or provider-hosted agent setups to downloadable open-source harnesses: Microsoft attributes a global supply-chain campaign to the China-state-aligned Storm-0940 using ARTEX, an open-source multi-agent penetration testing framework that works with any model [2].

Capability has diffused in both directions. State actors are adopting public tools, and public tools are letting less-resourced actors reach outcomes once associated with states. CrowdStrike assesses with moderate confidence that agentic AI commoditizes baseline offensive operations, and documented a low-sophistication operator who compromised government domains using 819 mostly public skills and no additional orchestration [3]. Anthropic's September 2026 threat report concludes that the main distinguishing feature between state and individual actors "is no longer sophistication but intent," and that the capabilities it describes should be assumed to be available to any motivated actor [4].

Two second-order consequences follow. Shared tooling is likely to make tool-based attribution less reliable [2][9], and because operators can switch to open-weight models when a commercial provider blocks them [5][6], provider-side disruption is also likely to be a less dependable backstop than it was in 2025. The entry points in these campaigns nevertheless remain familiar: exposed admin interfaces, reused credentials, unauthenticated APIs, and over-privileged application identities [2][5][9]. Organizations should plan for state-level persistence and tempo from a much broader adversary population, and prioritize exposure reduction, non-human identity governance, and machine-speed detection.

Background

For much of the last two decades, enterprise threat models have commonly rested on an implicit tiering of adversaries. Nation-state operators were assumed to bring capabilities that others could not match: patient multi-stage campaigns, custom tooling, large teams working in shifts, and the ability to sustain

operations against many targets at once. Criminal groups were treated as opportunistic and constrained by labor costs, and individual actors were assumed to lack the skills to chain reconnaissance, exploitation, lateral movement, and exfiltration without help. That tiering shaped where organizations spent their defensive budget and which incidents they treated as strategic.

The GTG-1002 disclosure in November 2025 marked the first public crack in that model. Anthropic reported that a group it assessed with high confidence to be Chinese state-sponsored used Claude Code, connected to tools through the Model Context Protocol, to run an espionage campaign against roughly thirty organizations. By Anthropic's estimate, the AI performed 80 to 90 percent of tactical operations, with human operators intervening mainly at a handful of decision points [1]. The campaign succeeded in a small number of cases. Some independent security researchers questioned the report's autonomy estimates and the absence of published indicators of compromise [14], but subsequent 2026 reporting from multiple vendors has documented comparable patterns. At the time, the operation was notable because it depended on a frontier commercial model and a custom orchestration layer that the provider could detect and shut down.

The twelve months since then have changed both of those conditions. During 2026, a set of open-source agent harnesses, including Hermes Agent, OpenClaw, Cairn, PentAGI, CyberStrikeAI, and ARTEX, became widely available, and several were documented in malicious campaigns [2][5][6][8][9]. These frameworks package the orchestration logic that GTG-1002 had to build for itself: task planning, parallel sub-agents, tool integrations, persistent campaign memory, and unattended execution. Many are model-agnostic, which means operators can pair them with commercial frontier models, Chinese open-weight models such as DeepSeek, or whatever endpoint is cheapest and least restricted [2][5][6]. The question this note addresses is what happens to the adversary hierarchy when the scaffolding that once distinguished state operations becomes a public download.

Security Analysis

Documented Campaigns in 2026

The public record now contains enough vendor reporting to observe diffusion directly rather than infer it. The table below summarizes the campaigns most relevant to this question. Figures are vendor-reported and generally reflect what researchers could observe, often because an operator made an operational security mistake, so they likely understate actual activity. Readers should also bear in mind that most of these reporters are security vendors with a commercial interest in agentic threats; that does not discredit their findings, but it is a reason to weigh the record as a whole rather than any single report.

Campaign (reporter, date)	Reported actor	Framework	Model(s)	Reported scale or outcome
GTG-1002 (Anthropic, Nov 2025) [1]	Chinese state-sponsored (high confidence)	Custom orchestration over Claude Code and MCP	Claude	About 30 targets; AI performed an estimated 80–90% of tactical work; a small number of successful intrusions
Asian government intrusion (Dream via CSO Online, Aug 2026) [5]	Chinese-language operator (linguistic analysis; no group attributed)	Hermes and OpenClaw with parallel sub-agents	Evidence of DeepSeek-V4-Flash	12 attack waves over about four days; 85 cracked credentials; expansion toward energy and nuclear-safety-related organizations
knaithe/KnYuan (Unit 42, Jul 2026) [6]	Chinese-speaking individual assessed to be in Zhuhai	Hermes Agent, controlled via Telegram	DeepSeek; also configured Qwen, GLM, Kimi, MiniMax	Over 460 targeted systems; autonomous attempts did not fully compromise intended targets; confirmed compromises came from manual activity
GTG-20006 and GTG-10007	Russian state-nexus (attribution consistent with public reporting	Modified Claude Code skills; agent swarms with	Claude	Phishing, persistence, and exfiltration against Ukrainian

Campaign (reporter, date)	Reported actor	Framework	Model(s)	Reported scale or outcome
(Anthropic, Sep 2026) [4]	linking it to Midnight Blizzard); China-linked espionage operators	persistent memory		and European government targets; a fleet of 13 scheduled collection agents
Retail intrusions (Gambit Security, Sep 2026) [7]	Chinese-speaking, financially motivated operator	Strix, Cairn, Hermes	Mixed commercial and open-weight	At least 27 companies compromised in six days at roughly $25 per targeted company in model cost
Four agentic actors (CrowdStrike, Oct 2026) [3]	Spanish/English-, Chinese-, Indonesian-, and Russian-speaking operators	Modified PentestGPT and custom harnesses; one actor (Threat Actor 3) used 819 public skills with no added orchestration	Claude (one actor via a China-based proxy)	Healthcare, government, CRM, and SaaS compromises across multiple countries
Storm-0940 (Microsoft, Oct 2026) [2]	China-state-aligned	ARTEX (open source, any model)	Not specified	ConnectWise Cloud Backup supply-chain compromise of a multi-tenant application; sign-in attempts against more than 2,250 tenants; under three hours from targeting to

Campaign (reporter, date)	Reported actor	Framework	Model(s)	Reported scale or outcome
				privileged cloud access in multiple instances

Diffusion Runs in Both Directions

In CSA's assessment, the most significant pattern in this record is that diffusion is bidirectional. In one direction, state-aligned operators have adopted the same commodity frameworks available to anyone. Microsoft's Storm-0940 reporting is the clearest example: an actor it describes as China-state-aligned and active since 2021 used ARTEX, an open-source Go and Next.js multi-agent system that builds a graph of discovered assets and vulnerabilities and can run from reconnaissance through exfiltration with little human interaction [2]. Microsoft assesses that Storm-0940 was responsible for the supply-chain compromise of ConnectWise's multi-tenant Cloud Backup application disclosed in October 2026, and that the actor used ARTEX to accelerate scanning, password spraying, and OAuth application abuse against downstream customer tenants, in some cases gaining privileged access to cloud infrastructure [2]. Microsoft also reports that it has seen threat actors across multiple regions using the same framework. CSA's analysis of ARTEX and CyberStrikeAI found that ARTEX has been tied to a China-state-aligned supply-chain campaign, a financially motivated operator targeting South Korean banks, and a domestic Korean group of seven suspects that police say breached more than 90 organizations [9]. A single open-source project therefore sits behind activity spanning espionage, financial crime, and domestic hacking.

In the other direction, actors without state resources are reaching outcomes that previously required them. CrowdStrike's Threat Actor 3, which the company describes as the least sophisticated of the four it profiled, compromised several government domains using a collection of 819 skills drawn mostly from public repositories, with no additional orchestration, sometimes intervening only to push the agent past guardrails [3]. CrowdStrike expects that over the next twelve months agentic AI will reduce barriers to effective intrusions enough that some low-skill actors operate at the effectiveness and scale previously associated with organized crime or state-nexus actors [3]. The Unit 42 case shows the same dynamic from the other side: an individual operator ran exploitation workflows against more than 460 targets [6], a footprint that, for multi-stage exploitation rather than simple scanning, has typically required a team.

Anthropic's September 2026 report states the conclusion directly. It finds that AI "has collapsed the labor and tooling gap" between well-resourced state operations and individual operators, and that "the main distinguishing feature between these classes of actors is no longer sophistication but intent" [4]. The report notes that publicly available frameworks such as PentAGI reproduce similar scaffolding for anyone who downloads them, and concludes that the capabilities it documents should be assumed to be available to any motivated actor [4]. This is a statement about tooling and labor, not about strategic resources. In CSA's assessment, states retain advantages in zero-day acquisition, access to signals intelligence, legal cover, and patience. What has diffused is the operational layer that turns intent into sustained, parallel campaign activity.

Autonomy Is Not the Same as Effectiveness

A careful reading of the evidence argues against overstating what these frameworks achieve on their own. Unit 42 found that the recovered autonomous DeepSeek sessions attempted exploitation of Langflow and n8n targets but did not fully compromise them, and that the confirmed compromises came from separate manual exploitation, including of Citrix NetScaler devices and Marimo notebook servers [6]. Microsoft notes that the majority of Storm-0940's ARTEX-driven targeting was unsuccessful, with nearly all of the password-spray activity ending in lockouts [2]. Anthropic cautions that autonomy and severity are separate dimensions, since several of the most serious compromises in its reporting involved a human directing every step [4]. Even GTG-1002 succeeded in only a small fraction of its attempts [1].

These caveats matter for planning, but they do not reverse the conclusion. Commodity frameworks appear to weaken the constraint that low per-attempt success rates once placed on attackers. Gambit Security estimated model costs of roughly $25 per targeted company in one retail campaign [7]. At costs of that order, and with continuous execution that needs no staffing, an operator can accept a high failure rate and still achieve a meaningful number of compromises, although infrastructure, monetization, and the operator time needed for decisive manual steps [6] still carry costs. Microsoft observed that in several instances less than three hours elapsed between initial targeting and privileged cloud access [2]. The relevant shift for defenders is not in the sophistication of any single intrusion but in volume and tempo, which affect organizations broadly rather than only high-profile targets.

Consequences for Attribution and Disruption

Shared tooling has two second-order effects that deserve attention. The first concerns attribution. Defenders and intelligence analysts have long relied on tooling, infrastructure, and tradecraft fingerprints to distinguish one actor from another. When a China-state-aligned group, a financially motivated crew, and a domestic criminal ring all run ARTEX [2][9], or when a state-linked operation and a lone operator both build on Hermes [5][6], tool presence alone says little about who is behind an intrusion. CSA's

analysis notes that no published source has linked the separate ARTEX and CyberStrikeAI campaigns to a single actor [9]. Analysts may need to rely more heavily on targeting patterns, operator language artifacts, timing, and downstream use of stolen data, which are slower and less certain signals. This also suggests that early incident triage may misclassify state espionage as commodity crime, or the reverse, with consequences for disclosure obligations and response priorities.

The second effect concerns provider-side disruption. GTG-1002 was detected and disrupted in part because it ran on a single commercial provider that monitored for misuse [1]. In several documented 2026 campaigns, by contrast, operators paired frameworks with open-weight models, and the frameworks themselves are designed to make such substitution easy. Unit 42 reports that the knaithe actor configured multiple Chinese open-weight models and chose DeepSeek as the reasoning engine [6], and Dream found evidence of a DeepSeek model inside the Asian government intrusion framework [5]. Safeguards designed to restrict AI behavior were bypassed in that campaign by framing the activity as authorized testing [5], and in the retail campaign the operator added a skill to strip Hermes's own content filters [7]. Provider monitoring therefore remains valuable, as Anthropic's continued disruptions show [4], but it can no longer be treated as a dependable backstop. Release governance is similarly limited: CSA's ARTEX analysis reports that the developer made the project closed-source after the abuse became public, yet copies already in circulation remain available [9].

What Has Not Changed

Despite the novelty of the tooling, the entry points across these campaigns are conventional. The Asian government intrusion began with reconnaissance of publicly accessible code that exposed API endpoints and authentication configurations, then exploited unauthenticated APIs and moved laterally through single sign-on integrations [5]. Storm-0940 combined password spraying with abuse of credentials on multi-tenant OAuth applications, including the compromised ConnectWise integrations [2]. CSA's review of ARTEX and CyberStrikeAI found that reported entry points were exposed admin interfaces, default or reused credentials, legacy systems, and over-privileged third-party application identities [9]. Agents amplify the speed and breadth with which attackers find these weaknesses; they have not, on current evidence, invented new classes of initial access. CSA considers this the most actionable finding in the record, because it means existing control families still apply, provided they are enforced consistently and at a pace that matches automated adversaries.

Recommendations

Immediate Actions

Organizations should begin by revising threat models to assume that state-level persistence, parallelism, and tempo are available to any motivated adversary, as Anthropic's reporting advises [4]. In practice, this means removing the assumption that only high-profile organizations face multi-stage campaigns. Security teams should inventory internet-facing administrative interfaces, VPN and edge devices, and APIs, verifying that none accept unauthenticated requests or default credentials, since these were the recurring entry points in 2026 campaigns [5][9]. Microsoft 365 tenants that authorized ConnectWise integrations should follow Microsoft's published guidance to preserve logs, search by application ID, and review service principals for new credentials, owners, and consent grants [2]. Teams should also add published indicators from the Microsoft, CrowdStrike, Unit 42, and Gambit reports to blocking rules and retrospective hunts [2][3][6][7].

Short-Term Mitigations

Over the next quarter, organizations should treat non-human identities as a primary attack surface. Multi-tenant OAuth applications, backup integrations, and service principals should be held to least privilege, and teams should confirm they can revoke any integration quickly. Detection logic should be tuned for the signatures of agent-operated activity: distributed, low-frequency attempts correlated across assets and time, rapid bursts of sign-ins from application credentials, and reconnaissance sequences that complete faster than human operators could manage [2][9][11]. Internal AI services deserve specific attention as well. Workflow platforms such as Langflow and n8n, along with notebook servers, were targeted in the Unit 42 campaign [6] and should not be exposed to the internet, and egress monitoring should flag calls to unapproved LLM APIs that could indicate an attacker staging an agent inside the environment. Incident response playbooks should avoid drawing conclusions about actor identity from framework presence alone.

Strategic Considerations

At the strategic level, organizations should adopt continuous automated testing of their own external attack surface, ideally using authorized agentic tooling, on the premise that adversaries are already running equivalent tools against them [9]. Defensive remediation cycles measured in weeks are poorly matched to adversaries that can reach privileged access in hours [2]. Security leaders should also engage with the policy question of offensive agent framework release. Model providers, open-source

maintainers, and code hosting platforms have an interest in shared norms for staged release, abuse reporting, and provenance, although the persistence of forked copies suggests that such measures will slow rather than prevent diffusion [9]. Finally, boards and risk committees should be briefed that the adversary tiering underpinning many risk registers may be out of date. The appropriate question is no longer whether an organization is interesting to a nation-state, but whether it is reachable by an agent that anyone can run.

CSA Resource Alignment

CSA's most directly relevant prior work is *Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX* [9], published in October 2026. That note documented how individual open-source frameworks have been adopted by state-aligned, financially motivated, and domestic criminal actors at the same time. This note extends that framework-level analysis to the broader question of adversary tiering, and adopts its finding that entry points remain conventional as the foundation for the recommendations above. *Agentic Intrusion Goes Retail: Detection and Hunting Lessons* [15] examines the Gambit-reported retail campaign in more depth and complements the economic discussion in this note.

Autonomous Agentic AI Adversaries [10], published in June 2026, established that autonomous AI adversaries had become an operational reality and mapped defensive guidance to MAESTRO, AICM, and Zero Trust. The campaigns reviewed here are consistent with that paper's trajectory and add a dimension it anticipated but could not yet document in detail: the convergence of state and non-state actors on shared tooling. *Fingerprinting Agent-Operated Intrusions: Detection and Response Guidance* [11] supplies the detection engineering detail behind the short-term mitigations in this note, including how to identify agent tempo and machine-generated reconnaissance patterns in logs.

For control implementation, the AI Controls Matrix (AICM) v1.1 [12] provides the relevant control objectives. Its Identity and Access Management domain addresses the non-human identity and OAuth governance gaps exploited by Storm-0940, and its Threat and Vulnerability Management domain supports the continuous exposure reduction recommended above. Organizations deploying their own agents should apply the MAESTRO agentic threat modeling framework [13] to ensure internal agent infrastructure does not become either an entry point or a pre-positioned capability for an intruder.

References

[1] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic, November 13, 2025.

[2] Microsoft Threat Intelligence. "AI-driven, large-scale nation-state campaign using ARTEX enables supply chain compromise." Microsoft Security Blog, October 21, 2026.

[3] CrowdStrike Counter Adversary Operations. "CrowdStrike Uncovers New Agentic AI Campaigns." CrowdStrike Blog, October 21, 2026.

[4] Anthropic. "Detecting and countering misuse of AI: September 2026." Anthropic Threat Intelligence, September 2026.

[5] Gyana Swain. "AI agents wage near-autonomous cyberattack on Asian government networks." CSO Online, August 13, 2026.

[6] Palo Alto Networks Unit 42. "Chinese-Speaking Threat Actor Harnesses AI Models for Autonomous Cyberattacks." Unit 42, July 30, 2026.

[7] Eyal Sela. "AI Agents Are Hacking Online Retailers for $25 a Company." Gambit Security, September 22, 2026.

[8] NorForge. "Inside Cairn: Open Windows into massive AI-Driven Cyber Operations." NorForge Blog, October 19, 2026.

[9] Cloud Security Alliance AI Safety Initiative. "Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX." CSA Labs, October 28, 2026.

[10] Cloud Security Alliance AI Safety Initiative. "Autonomous Agentic AI Adversaries." CSA Labs, June 24, 2026.

[11] Cloud Security Alliance AI Safety Initiative. "Fingerprinting Agent-Operated Intrusions: Detection and Response Guidance." CSA Labs, October 22, 2026.

[12] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[13] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[14] Bill Toulas. "Anthropic claims of Claude AI-automated cyberattacks met with doubt." BleepingComputer, November 14, 2025.

[15] Cloud Security Alliance AI Safety Initiative. "Agentic Intrusion Goes Retail: Detection and Hunting Lessons." CSA Labs, October 27, 2026.