CSAI Foundation | Frontier Ready Cybersecurity

Frontier Ready Daily

Issue 92 · Member edition

2026-11-26



CSAI FOUNDATION INITIATIVE

Frontier Ready.

Machine-speed agentic cybersecurity

cloud security alliance®

CSAI

© 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.

In this issue

The lead is Switzerland's NCSC confirming its first AI-agent intrusions: every attack it knows of succeeded, and it now tells organisations to install critical updates on internet-facing systems within 24 hours. It is followed by a measured 51-second median from publishing an image with a cloud key to that key being used, a Drupal flaw found by an AI scanner and exploited about a day after its fix, the counters that leak from agentic attack frameworks into victim logs, Google's 1 February 2027 deadline for moving Workspace agents onto governed access paths, and Anthropic opening cross-session misuse monitoring under zero data retention to public beta. We examined the 1Password GitHub-account breach and left it out because nothing in the disclosure involves agentic tooling or a machine-speed interval; CSA's research note "1Password Backend Breach: When the Secrets Broker Is the Target" covers it.

Switzerland's NCSC says every AI-agent attack it knows of compromised the victim, and sets a 24-hour limit for critical patches on internet-facing systems

Category: security_operating_model **Significance:** major **Confidence:** VERBATIM (PROVIDER) for "In all cyberattacks known to the NCSC, the use of AI agents led to the successful compromise of the victim's IT systems" and for the 24-hour patching recommendation; SELF-REPORTED (PROVIDER METRIC) for internet-accessible AI agents having "almost doubled since October 2026", given with no baseline or method; NO PROVIDER CLAIM on how AI-agent involvement was established, how many organisations were hit, or which sectors; CHARACTERIZATION (CSA) that the 24-hour figure is a usable external benchmark for patch SLAs **Source:** National Cyber Security Centre NCSC – Cyberattacks involving AI agents: First confirmed cases in Switzerland

What changed. On 24 November 2026 the Swiss NCSC reported its first confirmed attacks carried out "with the assistance of an AI agent" against Swiss organisations, including critical-infrastructure operators in at least two sectors. It said every such attack it knows of succeeded, named agentic offensive frameworks (ARTEX and CyberStrikeAI as examples) whose internet-accessible instances have almost doubled since October, and told organisations to install critical updates on internet-reachable systems "no later than 24 hours after release."

Why it reaches you. The initial access paths the NCSC names are default or weak passwords and known flaws on internet-facing systems such as webmail and VPN gateways. These are the same paths that monthly patch cycles and exception-based MFA leave open. This is a national authority, not a vendor,

putting a 24-hour number on critical patching of exposed systems. Most enterprise patch SLAs for critical findings are measured in days or weeks, and a regulator or auditor can now point to that number. The NCSC shares detection indicators only with critical-infrastructure operators. Other organisations get the guidance but not the indicators.

What to do. Vulnerability management lead: validate. Measure your actual time-to-install for the last five critical updates on internet-facing systems against the 24-hour figure, and list every internet-reachable service still exempt from MFA. Take both gaps to the CISO as a funded exception or a remediation plan.

CSA coverage. Machine-Speed Attacks: Cloud Defense at the Inflection Point

Cloud keys left in public container images were first used a median 51 seconds after publication, and every planted AWS key was used within 15 minutes

Category: machine_speed **Significance:** major **Confidence:** VERBATIM (PROVIDER) for the study's intervals (fastest 7.86 seconds, median 51 seconds, 93% within five minutes, all within 15 minutes) and for "publication and compromise are, for practical purposes, the same event"; CHARACTERIZATION (CSA) that the result applies to AI coding-agent state files baked into images; LINK ONLY – VERIFY AT SOURCE for the peer-reviewed paper's full method **Source:** Help Net Security – Bots exploited leaked AWS key from Docker image in under 8 seconds

What changed. Stony Brook University researchers published 300 decoy nginx images containing AWS, OpenAI and Anthropic keys across eight registries over four months. In the first round of 100 Docker Hub images, the fastest key use came 7.86 seconds after the push and the median 51 seconds. Every AWS key was used within 15 minutes, and the keys drew more than 140,000 API calls from more than 5,000 IP addresses. The bots never ran the containers. They read a public catalog of newly published images, guessed which layers held credentials, and pulled only those layers. AI-provider keys followed a "validate-wait-exploit" pattern: one OpenAI key sat idle for about a month and was then drained in under 24 hours, which the authors suspect reflects resale. Keys with restricted permissions received no token-consuming calls.

Why it reaches you. This puts a number on the exposure window for any credential in a published image, and the window is shorter than any human alert-and-revoke process. Docker Hub's own pull counter lags actual downloads by five to six minutes, so an owner watching downloads learns of the exposure after the key has been used. The same failure now runs through AI coding agents. GitGuardian separately counted 613 distinct secrets, 99 of them valid, in Claude Code settings files shipped inside public Docker Hub images, because `COPY . .` copies the working tree. We covered that file's secret leakage in the 7 November issue. Model-provider keys are part of what gets taken, and they are sold on.

What to do. Platform engineering lead: escalate. Make secret scanning of image layers a blocking build step before any push to a public or partner-readable registry. Scope every model-provider and cloud key placed in a build context to the minimum permissions, since restricted keys were the only ones attackers left alone. Treat any key found in a published image as already used: revoke it, don't delete the tag.

CSA coverage. New — no prior CSA coverage.

Drupal's installer flaw was found by an AI scanner, exploited about a day after the fix, and has no patch for end-of-life branches

Category: vuln_storm **Significance:** major **Confidence:** VERBATIM (PROVIDER) for the exploitation report "on live sites as early as 2026-11-19 19:00 UTC", for the credit to Anthropic's OSS Scanner, and for the end-of-life statement; NO PROVIDER CLAIM on the number of sites exploited or patched, or on whether attackers used AI; CHARACTERIZATION (CSA) that the fix-to-exploit interval was roughly one day **Source:** Drupal.org — Drupal core - Highly critical - Remote code execution - SA-CORE-2026-017

What changed. Drupal published SA-CORE-2026-017 (CVE-2026-132165, risk 20/25) on 18 November 2026. The flaw lets an unauthenticated attacker use `install.php` to try to reinstall a live site, which can lead to takeover or remote code execution. Anthropic's OSS Scanner found it. The advisory was later updated to report exploitation on live sites from 19 November at 19:00 UTC. Searchlight Cyber published a working exploitation write-up on 24 November, and CISA added the flaw to its Known Exploited Vulnerabilities catalog with a 27 November federal deadline. Fixes exist only for 10.6.22, 11.3.22 and 11.4.13. Drupal 11.2.x and below, 10.5.x and below, and Drupal 8 and 9 get none.

Why it reaches you. The bottleneck here is the gap between discovery and the sites able to take the fix, not the discovery itself. An AI scanner found the flaw, attackers used it within about a day of the fix, and every site on an end-of-life branch is left with a workaround (blocking the installer path) instead of a patch. Marketing sites, intranets and partner portals often run Drupal outside central patch management, and they are the most likely to be on an unsupported branch.

What to do. Web application owner: escalate. Inventory every Drupal instance, including agency- and marketing-run ones, and upgrade to a fixed release. On any end-of-life branch, block `/core/install.php` at the web server today and schedule the branch upgrade. Check web logs for installer requests since 19 November.

CSA coverage. When Discovery Outruns Remediation: Open-Source Disclosure Strain

Agentic attack frameworks leak incrementing intent IDs into cloud resource names and User-Agents, which defenders can use to tie rotating IPs and separate clouds to one attack run

Category: machine_speed **Significance:** notable **Confidence:** SELF-REPORTED (PROVIDER METRIC) for Wiz's counts (over a hundred harness repositories reviewed, hundreds of distinct intent IDs catalogued, 36 AWS-side and 34 Azure-side IDs in one sequence); VERBATIM (PROVIDER) for the framework code paths that place the ID in prompts and working directories; CHARACTERIZATION (CSA) that this gives SOC teams a correlation key that survives proxy rotation; LIVE TEST REQUIRED for false-positive rates of the suggested detections in a given estate **Vendor interest:** Wiz sells cloud detection and response, so detection guidance built on its own telemetry is commercially favourable to it. **Source:** Wiz – Inside Agentic Offensive Frameworks: Intent IDs

What changed. On 25 November 2026 Wiz showed that ARTEX, Cairn and their forks number each attack sub-goal ("intent") and place that number in the model's prompt and working directory. The model then reuses it when naming things in the victim's environment: User-Agents (`artex-verify-i89xx`), GitHub workflow files, test accounts, IAM role session names, S3 buckets, SSM comments, KMS aliases and Entra app registrations. These names survive even in error messages from denied API calls. ARTEX's counter never resets per installation, so climbing IDs tie activity together across rotating IPs. In one intrusion, a single sequence linked an AWS-side and an Azure-side intrusion run through identities that had nothing in common.

Why it reaches you. Agentic operators rotate proxies, so IP-based clustering in the SIEM breaks. This gives cloud audit logs (CloudTrail, Entra audit, CI workflow history) a correlation key that does not depend on infrastructure. It can also show whether an installation is focused on you or spraying many targets at once. Wiz is explicit about the limits: Cairn restarts at `i001` for every project, new installations collide in the low numbers, and the signal corroborates attribution but does not prove it.

What to do. Detection engineering lead: validate. Add hunts for `i<N>` patterns in cloud resource names, session names, tags, User-Agents and command lines, alerting only when paired with an anchor such as an `artex-`, `pentest-`, `probe` or `canary` prefix or a `tasks/<task>/i<N>` path. Run them back over 60 days of cloud audit logs before enabling alerts.

CSA coverage. Agentic Intrusions in the Wild: Detection and Response After Chainalysis

Google will require AI agents to reach Workspace data through its MCP servers or CLI from 1 February 2027, with dedicated client IDs and user confirmation for sensitive actions

Category: agentic_surface **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for "AI agents must use Workspace MCP servers or the Workspace CLI rather than existing REST APIs. Technical enforcement begins on February 1, 2027" and for the confirmation, client-ID and revocation controls; NO PROVIDER CLAIM on what happens after the deadline to agents that use browser automation or other paths; LIVE TEST REQUIRED for whether the confirmation step holds when a prompt-injected agent asks for it **Source:** Google Cloud Blog – Built for trust and control: Powering the next wave of AI agents across Google Workspace and Google Cloud

What changed. On 24 November 2026 Google made its Workspace MCP servers and CLI generally available and published a Workspace AI Tools Addendum. Under the addendum, agents must stop using the REST APIs by 1 February 2027. Conventional apps keep REST access, and the MCP servers reject non-agent OAuth clients. Agents must use dedicated OAuth client IDs with least-privilege scopes, and quota billing begins on 1 April 2027. The MCP servers and CLI require explicit user confirmation before an agent sends email, changes Drive sharing or makes bulk edits. Admins can view, audit and revoke agents in the admin console. The addendum also bars permanent raw storage, training on Workspace data and API scraping.

Why it reaches you. Third-party and in-house agents that read corporate mail and documents through a generic OAuth grant become distinguishable as agents for the first time, which gives the identity team something to inventory and revoke. The deadline also means every agent integration against Workspace has to be re-plumbed within about ten weeks, and integrations no one owns will be found when they break. The confirmation gate covers sending, sharing and bulk edits. It does not cover reading, which is where prompt-injected exfiltration starts.

What to do. SaaS identity owner: validate. Pull the list of OAuth grants with Workspace mail, Drive and Calendar scopes and identify which are agents. Require each owner to confirm a migration plan before 1 February 2027, and revoke any grant with no owner.

CSA coverage. Who Did That? Agent Action Attribution as an Audit Obligation

Anthropic opens cross-session misuse monitoring with zero data retention to public beta, sending flagged cases to the customer's SIEM without prompts or outputs

Category: defender_models **Significance:** context **Confidence:** VERBATIM (PROVIDER) for "Only automated safety monitoring can access this data" and for Safety Cases that "do not include inputs and outputs"; NO PROVIDER CLAIM on detection rates, false-positive rates or eligibility criteria; LIVE TEST REQUIRED for whether multi-session agent misuse is actually flagged **Vendor interest:** Anthropic sells the Claude platform this monitoring runs on, so the case for adopting it comes from the party that profits from it. **Source:** Claude – Zero data retention with Enterprise Frontier Safeguards is now in beta

What changed. On 23 November 2026 Anthropic opened Enterprise Frontier Safeguards to public beta for eligible customers on its own platform, Amazon Bedrock and Google Cloud. Inputs and outputs are written to customer-owned S3 or Cloud Storage under customer keys, and only automated monitoring reads them. When it finds a risk across sessions, including agents exceeding what their users intended, it creates a "Safety Case" that the customer can read through storage or an API and feed into a SIEM.

Why it reaches you. Until now, zero-data-retention contracts screened each request alone, and attacks split across sessions passed through. That left enterprises choosing between data minimisation and provider-side abuse detection. This changes what a model-gateway or AI procurement review can ask a provider for. It is still a provider-run detector with no published efficacy, and no independent test exists yet.

What to do. AI governance lead: monitor. In the next model-provider contract review, ask whether cross-session misuse detection is available under your retention terms and whether its alerts reach your SIEM.

CSA coverage. Canary Evidence of Prompt Theft at AI Inference Providers

Watchlist

- **OpenAI reward-hacking postmortem – downstream response** – No change. *(opened 2026-08-27)*
- **VM/hypervisor containment hardening for cyber-capable agents** – No change. *(opened 2026-08-27)*
- **Claude Code Auto Mode prompt-injection ASR discrepancy** – No change. No Anthropic response to the reproduced attack success rate. GitGuardian reports that the Claude Code

lead recommended auto mode over saved permission rules as a secrets mitigation, but that does not address injection success rates. *(opened 2026-08-27)*

- **AI defensive-triage guardrail evasion** – No change. *(opened 2026-08-31)*
- **AI account session hijacking at scale** – A related signal, not a session-hijacking campaign. The Stony Brook study (item 2) shows OpenAI and Anthropic API keys taken from public images being validated, held for weeks and then drained, which the authors read as a resale market for AI-provider credentials. No provider has disclosed a session-hijacking campaign or shipped device-bound session tokens. *(opened 2026-08-31)*

Opened this issue

- **Workspace agent-access enforcement (1 February 2027)** – `agentic_surface`. Watching for Google to say what happens to agents that use REST, browser automation or screen reading after the deadline, for the time-limited approvals it says are coming, for independent tests of whether the user-confirmation gate holds against prompt-injected agents, and for Microsoft 365 or Slack to adopt comparable agent-only access paths.