CSAI Foundation | Frontier Ready Cybersecurity

Frontier Ready Daily

Issue 69 · Member edition

2026-11-03



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 item is a measured intrusion timeline. Kaspersky reports that once operators handed a compromised education network to the CyberStrikeAI agent, it set up persistence within two minutes, carried out lateral movement, credential dumping and a database export in under an hour, and reached domain administrator within three to four days of deployment. The other five items: coding agents that install deprecated, unpatchable MCP servers from model memory even after npm warns them; Google's account of stolen cloud AI access resold through "LLM Mixers," with bills reaching $1 million; a suspected agentic loop in a 3AM ransomware attack on a UK school estate; Red Hat shipping fixed kernels for CVE-2026-98432 27 days after the upstream fix while Ubuntu's fixes remain opt-in; and HackerOne platform data showing bug-bounty submissions growing faster than valid findings.

An AI agent reached persistence two minutes after deployment and domain administrator within three to four days in attacks on education networks

Category: machine_speed **Significance:** major **Confidence:** VERBATIM (PROVIDER) for Kaspersky's timeline: persistence and local reconnaissance "within 2 minutes," domain reconnaissance "over the next 5 hours," lateral movement, credential dumping and database export in "no more than one hour," and agent deployment to domain administrator compromise "within 3–4 days"; VERBATIM (PROVIDER) for the architecture, in which the model runs on the attacker's C2 server and the victim-side implant has no built-in AI; NO PROVIDER CLAIM on victim count, named countries, or which model drove the agent; CHARACTERIZATION (CSA) for the reading that the human-paced gaps, not the agent's steps, now set the timeline **Vendor interest:** Kaspersky sells endpoint detection and managed detection and response, so the detection-focused recommendations are commercially favourable to it. **Source:** Kaspersky Securelist – CyberStrikeAI in the wild: the use of AI in real-world attacks on organizations

What changed. On 2 November 2026 Kaspersky described intrusions at education organizations in several African and Latin American countries using CyberStrikeAI, an open-source penetration-testing framework written in Go. In the incident it reconstructed, an operator spent nearly a day on manual access through a compromised SQL Server and then paused for two weeks. Once the operator deployed the agent, it established persistence within two minutes and mapped the domain over five hours. After a pause of several days, it completed lateral movement over SMB and WMI, credential dumping and a

database export in under an hour. Kaspersky did not attribute the activity to a known group, but notes Chinese-language script comments and C2 restarts during UTC+8 business hours. Help Net Security reported the findings alongside CrowdStrike's ARTEX campaign, which this feed covered on 30 October.

Why it reaches you. The component classes are internet-facing database servers and Active Directory. The agent's techniques were conventional: scheduled-task persistence, LSASS dumping through comsvcs.dll, Kerberoasting, DCSync and WMIExec-style lateral movement. Kaspersky's own assessment is that AI "does not yet create fundamentally new detection challenges", and that what changes is speed. CSA reads this as the agent compressing each active phase to minutes or hours, so the long gaps in the timeline are the operator's choices. A SOC that relies on catching a slow human operator during discovery will see that phase finish in a single shift.

What to do. Security operations owner: validate. Confirm that alerts for shells spawned by `sqlservr.exe`, comsvcs.dll MiniDump use, bursts of LDAP and TGS requests, and remote scheduled-task creation reach a responder within the hour rather than the next business day. Replay this sequence in a purple-team exercise and record the time from first alert to containment. Urgency: validate.

CSA coverage. Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX; see also Agentic Intrusion Goes Retail: Detection and Hunting Lessons.

Coding agents keep installing archived MCP reference servers that their maintainers will not patch, even after npm marks them deprecated

Category: agentic_surface **Significance:** major **Confidence:** SELF-REPORTED (PROVIDER METRIC) for Glow's finding that in 98% of studied sessions where a deprecated server was installed or suggested, the agent supplied the package name itself, and for the 17.3 million npm downloads since July 2025; VERBATIM (PROVIDER) for the maintainers closing Glow's Puppeteer-server remote code execution report as out of scope because the server was archived in May 2025; NO PROVIDER CLAIM of exploitation in the wild; CHARACTERIZATION (CSA) for the reading that model memory has become an unmanaged package source **Vendor interest:** Glow sells security tooling for AI agents, including the kind of pre-execution controls its research recommends, so the framing is commercially favourable to it. **Source:** Help Net Security – AI coding agents are installing MCP servers that will never be patched

What changed. On 3 November 2026 Glow Labs published research it calls MCPhantom, based on real coding-agent session transcripts from Claude Code, Cursor, Codex CLI and Claude Cowork running Anthropic, OpenAI and xAI models. In sessions where a deprecated `@modelcontextprotocol` reference server was installed or suggested, the agent supplied the package name from its training data 98% of the time, and the user asked for it by name only 2% of the time. In one GitHub case the agent

chose the archived reference server over GitHub's own maintained server. In another, the agent looked up the package on npm, received "This package is deprecated," and installed it anyway without telling the user. Glow found four high-severity flaws in the archived servers, including prompt-injection-driven code execution in the Puppeteer server, and none will be fixed. Npm's removal rules do not cover these packages. The GitHub server alone had 185,239 downloads last week and six owners.

Why it reaches you. The component class is the developer endpoint and its agent runtime, where `npx` fetches and runs an MCP server at the agent's request. Glow notes that EDR sees a trusted Node process doing routine work and SCA sees no manifest. The flaws carry no CVEs, so vulnerability feeds will not flag them, and software inventory never records the install. The deprecation warning, the only control in the chain that fired, was ignored by the agent that received it.

What to do. Endpoint and developer platform security owner: validate. Search developer machines and repositories for MCP configuration files that reference deprecated `@modelcontextprotocol` servers, and switch to the service owner's maintained server where one exists. Approve one server per service and make npm deprecation status a blocking criterion. Where your agent harness supports pre-execution hooks, use them to deny installs of unapproved MCP packages. Urgency: validate.

CSA coverage. AI Coding Agents as Attack Surface: MCP, Poisoning, and Miasma.

Stolen cloud AI access is resold through "LLM Mixers," and Google says single incidents exceeded $100,000 a day and reached $1 million before access was cut

Category: agentic_surface **Significance:** major **Confidence:** SELF-REPORTED (PROVIDER METRIC) for Mandiant's figures: more than $100,000 in a single day, cumulative charges up to $1 million, more than 500,000 model calls in under eight hours by one actor, and nearly $1 million in token costs passed to one victim by the Jet Night resale service; VERBATIM (PROVIDER) for CVE-2026-40217 in the LiteLLM gateway, which let standard virtual keys be elevated to `PROXY_ADMIN`; NO PROVIDER CLAIM on incident count, which is given only as "multiple incidents in 2026," or on time to detection; CHARACTERIZATION (CSA) for the reading that disabled AI services are not a safe default **Vendor interest:** Google sells cloud security operations and Mandiant incident-response services, and the post points to its own detection rule packs, so the threat framing is commercially favourable to it. **Source:** Google Cloud Threat Intelligence – LLMJacking: Monetizing Compromised Cloud Models

What changed. On 2 November 2026 Google Threat Intelligence Group and Mandiant described LLMjacking incidents across financial services, retail, technology and transportation in three regions. The post separates the break-in from the payoff. One group steals service-account keys, GitHub tokens or session tokens. "LLM Mixers" such as Jet Night and newapi.pro pool the stolen access and resell it, with Jet Night offering unlimited use for about $60 a month against a legitimate rate of $100 to $250.

Victims' AI services were often switched off when they were compromised, and attackers used the stolen keys to switch them on. Attackers tested keys with `max_tokens=1` requests to stay below volume alerts. They also used a fake portal to collect active tokens, which let them bypass MFA. This feed covered Wiz's 30-minute first-use figure for October's platform breaches on 1 November; Google adds the resale market, the gateway privilege-escalation flaw and the re-enablement path.

Why it reaches you. The component classes are the cloud model service (Bedrock, Azure AI and Vertex), the model gateway in front of it, and the service-account keys that reach both. An estate that never adopted a cloud AI service is still exposed if its IAM permits the service to be enabled. CSA notes that Google says Vertex data-access logs are off by default, so model invocations may not be logged until someone turns them on.

What to do. Cloud security owner: validate. Treat the enabling of any cloud LLM service, and any change to model invocation logging, as a high-severity event, and confirm the alert works in AWS, Azure and GCP. Deny AI-service permissions to principals that do not need them, set per-model spending caps, and patch LiteLLM or any other gateway you run against CVE-2026-40217. Urgency: validate.

CSA coverage. Harvesting Keys for Distillation: Attacks on AI Credential Platforms; see also Black October Reaches PaaS: Northflank, Secret Stores, and AI Key Resale.

A 3AM ransomware affiliate issued slightly varied commands at 40- to 100-second intervals across a UK school estate, which Bitdefender reads as a suspected agentic loop

Category: machine_speed **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for Bitdefender's observation of "a sequence of repeated commands that were slightly modified at each attempt" at regular intervals, as reported by BleepingComputer; VERBATIM (PROVIDER) for Bitdefender's caveat that "we cannot document anything about the language model itself"; NO PROVIDER CLAIM on the model, where it ran, or whether a human relayed instructions; CHARACTERIZATION (CSA) for the reading that command cadence is now a usable hunting signal **Vendor interest:** Bitdefender sells endpoint protection and managed detection, and its product was among those the attacker tried to disable, so the threat framing is commercially favourable to it. **Source:** BleepingComputer — UK school estate hit by 3AM ransomware in suspected AI-assisted attack

What changed. On 3 November 2026 BleepingComputer reported Bitdefender's investigation of a 3AM ransomware attack on an unnamed UK school estate between 10 and 19 September. The attacker came in from an unmanaged host. It extracted and decrypted credentials from three Veeam backup servers, dumped LSASS, created two backdoor domain admin accounts named "library" and "svc_bkpmon," and set up two-way trusts between three previously separate Active Directory domains. It

then encrypted about 37 of nearly 50 machines. Repeated, slightly modified commands arriving every 40 to 100 seconds led Bitdefender to suspect an automated agentic loop. It could not access the attacker's machine to confirm this.

Why it reaches you. The component classes are the backup server and the domain trust configuration. Backup consoles store privileged credentials and are frequently excluded from tight monitoring. Joining separate domains with new trusts turned one compromised estate into three. The evidence of AI assistance is behavioural and unconfirmed. Even so, a machine-regular cadence of retried, mutated commands is a pattern a human operator does not usually produce, so it can serve as a detection signal regardless of who or what is typing.

What to do. Identity and infrastructure security owner: validate. Alert on creation of new domain trusts and new domain-admin accounts, and treat either as an incident. Confirm that backup-server credentials are stored encrypted with keys held off the backup host, and that system-state backups cannot be deleted by a domain admin alone. Ask your detection team whether they can flag repeated near-identical commands at regular intervals from one session. Urgency: validate.

CSA coverage. Agentic Intrusion Goes Retail: Detection and Hunting Lessons.

Red Hat shipped fixed kernels for CVE-2026-98432 27 days after the upstream fix, while Ubuntu's fixes still do not install by default

Category: vuln_storm **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for Red Hat's 1 November advisories RHSA-2026:93412, 93413 and 93414 and the fixed RHEL 8.10, 9.8 and 10.2 kernel versions, as reported by The Hacker News; VERBATIM (PROVIDER) for Canonical's 2 November statement that fixes for 22.04, 24.04 and 26.04 LTS are in the "-proposed" pocket, which does "not install automatically," with no release date given; NO PROVIDER CLAIM of exploitation in the wild; CHARACTERIZATION (CSA) for the interval calculation from the 5 October upstream fix **Source:** The Hacker News – Red Hat Patches 9.8-Rated Linux TCP RCE Flaw While Ubuntu Releases Fixes for Testing

What changed. On 1 November 2026 Red Hat released fixed kernels for CVE-2026-98432, the remote IPv4 TCP use-after-free for which Anthropic reported an AI-written root exploit on 27 October. Red Hat first rated the flaw Moderate at 5.9 and has since raised it to Critical at 9.8. Its fix arrived 27 days after the 5 October upstream fix. On 2 November Canonical placed fixes for Ubuntu 22.04, 24.04 and 26.04 LTS in its opt-in `-proposed` channel. Cloud, FIPS and real-time variants are still in progress, and 20.04 and older releases are pending. As of this issue, a supported Ubuntu host that takes only regular updates is 29 days past the upstream fix with no installable kernel. On 1 November this feed reported that the interval had passed 25 days for both distributions.

Why it reaches you. The component class is the Linux server and container-host fleet, especially Ubuntu-based cloud images, which usually draw on the cloud kernel variant that is still in progress. Red Hat's initial Moderate rating shows that a vendor's own severity rating can lag the risk, so a patch SLA keyed to vendor severity would have put this flaw in a slower queue. Red Hat warns that its interim nftables filter applies only to the host where it is loaded and cannot stop container-to-container traffic on that host.

What to do. Infrastructure and platform security owner: escalate. Install and reboot into the fixed RHEL kernels now. For Ubuntu hosts facing untrusted networks, decide explicitly whether to take the `-proposed` kernel after testing or keep the IP-options filter in place, and record which applies to each fleet. Track the regular-channel release for cloud and FIPS variants. Urgency: escalate.

CSA coverage. CVE-2026-98432: Reported AI-Built Remote Root Exploit Precedes Distribution Fixes.

Bug-bounty submissions have grown 2.8 times since January 2025 against 1.9 times for valid reports, HackerOne says

Category: vuln_storm **Significance:** context **Confidence:** SELF-REPORTED (PROVIDER METRIC) for HackerOne's platform figures: submissions up 2.8 times and valid reports up 1.9 times since January 2025, valid AI-vulnerability reports up 222%, and payouts up 26% year on year in the first half of 2026; SELF-REPORTED (PROVIDER METRIC) for survey findings that about 70% of security leaders say validated findings reach their backlog faster than teams can address them; VERBATIM (PROVIDER) for the CEO's statement that the numbers "don't tell us how much of either increase comes from AI"; CHARACTERIZATION (CSA) for the implied fall in the valid share **Vendor interest:** HackerOne sells bug-bounty, triage and validation services, so evidence of rising submission volume and triage burden is commercially favourable to it. **Source:** Help Net Security – What human hackers still do better than AI, according to HackerOne's CEO

What changed. On 3 November 2026 HackerOne's CEO Kara Sprague gave Help Net Security platform figures that carry a denominator. Since January 2025 submissions have grown 2.8 times and valid reports 1.9 times, so by CSA's arithmetic valid findings have fallen from 100% to about 68% of their earlier share of submissions. She said she "can't give you a reliable split" between AI-generated and human reports. The same interview cites survey data that 58% of security leaders say business priorities delay fixes for critical and high-severity findings.

Why it reaches you. The component class is the external vulnerability intake pipeline: bounty programmes, disclosure inboxes and the triage staff behind them. If volume grows faster than validity, triage work grows faster than the number of real findings it produces. If 70% of leaders are right that

valid findings already outpace remediation, the extra triage effort also lands on a backlog that is not shrinking. No figure here is independently confirmed. A third-party count of submissions against validated reports across platforms would confirm it.

What to do. Vulnerability management owner: monitor. Track your own ratio of submissions to validated findings, and of validated findings to findings fixed, by quarter, so you have your own denominator rather than relying on a platform's. Urgency: monitor.

CSA coverage. When Discovery Outruns Remediation: Open-Source Disclosure Strain; see also VulnOps: Vulnerability Management in the Age of AI.

Watchlist

- **OpenAI reward-hacking postmortem — downstream response** — No change. No ruling in the Texas matter and no comparable eval-to-production escape disclosed by another lab. *(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 patch or public response. MCPhantom (item 2) shows Claude Code among the harnesses that install archived MCP servers, but it is a separate failure class. *(opened 2026-08-27)*
- **AI defensive-triage guardrail evasion** — No change. No new campaign this issue used refusals to blind security triage. *(opened 2026-08-31)*
- **AI account session hijacking at scale** — Partial movement. Google (item 3) documents a fake portal that collected active access tokens and bypassed MFA, plus stolen session tokens used to reach cloud AI services. That is session theft against cloud AI accounts, though not against AI provider consumer accounts specifically. No provider has shipped device-bound or short-lived session tokens. *(opened 2026-08-31)*

Opened this issue

- **Unmaintained agent components installed from model memory** *(agentic_surface)* — Glow reports that agents choose archived MCP reference servers from training data and ignore deprecation warnings (item 2). Watching for agent harnesses that refuse or warn on deprecated servers, for npm or the MCP project changing how archived packages are handled, and for the first exploitation of an unpatched reference server.