CSAI Foundation | Frontier Ready Cybersecurity

Frontier Ready Daily

Issue 89 · Member edition

2026-11-23



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 CERT-In's 20 November advisory. It says unauthorised payments went through legitimate approval workflows at service providers outside the financial sector, because the approvals ran inside the environment the attacker had already compromised. It is followed by a KVM flaw that let a guest read about 127 GB of a production cloud host's memory, the US intelligence chief's statement that emergency patching for the AI-found PHP2Shell flaws took down NSA's classified network, OpenAI's GPT-6.2 system card (exploit success up, and default-tier usefulness for allowed security work down), Google's pre-execution policy check on agent tool calls, and CrowdStrike's agreed purchase of Zscaler. We examined the 23 November round-up of Cairn sightings and left it out: its retail-campaign figures come from Gambit's report, which we covered on 25 September, and its NorForge figures we covered on 2 November. We also left out the Microsoft Entra Connect Health metadata exposure, because nothing in it involves agents or machine-speed operations.

CERT-In says approvals run inside a compromised service provider let unauthorised payments through, and that agentic attack frameworks are now active in every sector it hunts

Category: security_operating_model **Significance:** major **Confidence:** VERBATIM (PROVIDER) for CERT-In's statements that unauthorised transactions affected "technology service providers supporting government-linked payment processes and platforms handling vendor or partner payouts", that attackers "targeted business records and approval processes to carry out unauthorised transactions through otherwise legitimate systems", and that "multiple approvals within the same provider-controlled system may not provide sufficient protection if that environment is compromised"; VERBATIM (PROVIDER) for its threat-hunting finding of agentic offensive frameworks in activity against enterprises, service providers, government and academia; NO PROVIDER CLAIM for amounts lost, victim count, attribution, or whether AI was used in the payment-workflow incidents, which the advisory does not state; CHARACTERIZATION (CSA) for The Hacker News headline linking the payments to AI-assisted attacks, which goes further than the advisory text **Source:** CERT-In — CIAD-2026-0062: Abuse of Business Payment Workflows and Use of AI-Driven Agentic Offensive Frameworks across Sectors

What changed. On 20 November 2026 CERT-In issued a Critical advisory that widens its October and early-November warnings beyond banks. Attackers changed the business records and approval relationships that decide who gets paid, at technology service providers and payout platforms. The

payments then went through the normal payment process and passed the usual application and approval checks. In a separate section, CERT-In reports that its threat hunting found agentic offensive frameworks used against enterprises, technology providers, government and research institutions. It lists parallel multi-agent reconnaissance, exploits built quickly from patch analysis, and scripts rewritten when blocked. CERT-In's 15 October advisory had already reported that in one campaign agentic tooling cut the time from initial access to impact "from days to a few hours". The new advisory does not say whether the payment incidents themselves were AI-assisted.

Why it reaches you. The component class is the outsourced business-process platform: payroll, vendor-payout and disbursement systems run by a service provider, where both the record of who is owed and the approval of the payment live in the same tenant. If an attacker controls that environment, maker-checker controls inside it give no independent assurance, and each step looks authorised. Agentic tooling that tries many access paths in parallel makes it more likely that some provider in your supply chain is the environment that gets taken.

What to do. CFO's controls lead with the CISO and third-party risk owner: escalate. List every service provider that can change payee or beneficiary records or approve payments for you. Reconcile those records against an authoritative copy the provider cannot write to, and require out-of-band verification for changes to beneficiaries, payment instructions and approval chains. Write payment suspension, log access and shared-compromise notification into provider contracts. Urgency: escalate.

CSA coverage. ARTEX and Its Forks: Agentic Pentest Frameworks as Attack Infrastructure; see also Agentic Intrusions in the Wild: Detection and Response After Chainalysis.

A guest VM read about 127 GB of a production cloud host's memory through a KVM flaw on AMD hosts with nested virtualization, and most distribution kernels are still unfixed

Category: agentic_surface **Significance:** major **Confidence:** VERBATIM (PROVIDER) for Terra Security's and DigitalOcean's accounts: root inside a DigitalOcean Droplet read about 127 GB of host RAM read-only, a 32-byte write into a live host kernel page was shown on Terra's own bare-metal AMD server, the bug was introduced by the 2022 fix for CVE-2022-3344, and the fix shipped in stable 7.3.5 and long-term 6.18.64 and 6.12.118 on 17 November; VERBATIM (PROVIDER) for Terra's PatchSlop figure that 3.2% of kernel CVEs were introduced by the fix for another CVE; LINK ONLY – VERIFY AT SOURCE for The Hacker News' snapshot of Debian, Ubuntu and Red Hat fix status, the conflicting 8.8 and 5.6 severity scores, and VirtFusion's statement that nested virtualization is on by default in kvm_amd; NO

PROVIDER CLAIM for exploitation in the wild or for whether AI tools were used to find the flaw; CHARACTERIZATION (CSA) for the relevance to agent sandboxes **Source:** The Hacker News — Linux KVM Flaw CVE-2026-99312 Lets Guest VMs Access Host Memory on AMD Hosts

What changed. On 18 and 19 November 2026 Terra Security and DigitalOcean disclosed CVE-2026-99312. It is a stale page-table pointer in KVM's nested virtualization support for AMD processors. When an INIT signal stops a nested VM, KVM does not refresh the outer VM's page-table pointer, and the guest can then write its own page table over the freed page. Starting with root in its own Droplet, the guest mapped about 127 GB of the physical host's RAM. On a shared host, that memory includes other tenants' VMs. DigitalOcean live-patched its fleet. The upstream fix shipped on 17 November. Debian and Ubuntu still list their kernels as vulnerable, and Red Hat lists Enterprise Linux 9 and 10 as affected with the fix deferred. The flaw sat in the kernel for almost four years after a security fix introduced it.

Why it reaches you. The component class is the hypervisor boundary under agent sandboxes and multi-tenant compute. Execution sandboxes for code-running agents and evaluation harnesses often run as VMs, or as microVMs inside cloud VMs, which needs nested virtualization. An agent that gets root in its guest is the attacker position this flaw requires. Patching guest operating systems changes nothing: the fix lives in the host kernel and applies only after the host reboots.

What to do. Platform or infrastructure owner for self-hosted KVM: validate. On every AMD KVM host, read `/sys/module/kvm_amd/parameters/nested`. Either install a fixed kernel and reboot, or set `kvm_amd nested=0` where no tenant needs nested VMs. For agent sandboxes on cloud VMs, ask the provider whether its AMD hosts are patched and whether nested virtualization is exposed. Urgency: validate.

CSA coverage. New — no prior CSA coverage. For a related agent-sandbox escape, see Breaking the Agent's Egress Gate: Lessons from the Muse Sentinel Bypass.

The Director of National Intelligence says emergency patching for the AI-found PHP2Shell flaws left NSA's classified network unable to come back up as expected

Category: vuln_storm **Significance:** notable **Confidence:** VERBATIM (PROVIDER) for Director of National Intelligence Jay Clayton's statement, as reported by The Record and CBS News, that "the patching that has gone on following that has caused some of these systems to not come back up when you patch them as expected. That is what has happened at the NSA today"; VERBATIM (PROVIDER) for NSA's statement of "temporary degradation during scheduled routine maintenance"; LINK ONLY — VERIFY AT SOURCE for Nextgov's account of the 18 November House Homeland Security Committee testimony; LINK ONLY — VERIFY AT SOURCE for CBS News sources linking the outage specifically to

PHP2Shell patching and for the 21 October FBI/NSA/CISA instruction to patch or take systems offline; NO PROVIDER CLAIM linking the other November outages (Microsoft, FAA) to PHP patching, which the FAA traced to a router fault; CHARACTERIZATION (CSA) for treating patch-induced downtime as a cost of the vulnerability storm **Source:** The Record — Update: Disruptions in NSA's internal network caused by software patch, DNI says

What changed. On 17 November 2026 NSA Net, the agency's classified internal network, was degraded during what NSA called routine maintenance. The agency staggered staff arrivals at Fort Meade to keep tens of thousands of people from signing in at once. Clayton said the same day that the cause was patching for critical PHP vulnerabilities, which systems then failed to come back up from. CBS News sources tied it to PHP2Shell, the AI-found PHP flaws that FBI, NSA and CISA told organisations on 21 October to patch immediately or take offline. On 18 November officials told the House Homeland Security Committee that the November outage wave reflected maintenance and patching, not attacks. This is the first time a senior official has publicly blamed downtime at a major institution on emergency remediation of an AI-found flaw.

Why it reaches you. The component class is your emergency-change process for internet-facing web stacks, along with the older infrastructure behind them. When an AI-found flaw arrives with a "patch or disconnect" instruction, change windows shrink, regression testing gets cut, and systems that rarely restart get restarted all at once. The cost of the vulnerability storm then includes self-inflicted outages as well as breaches.

What to do. Head of IT operations with the vulnerability management owner: validate. For your PHP estate and other emergency-patched tiers, confirm that every system can be rolled back and has a tested restart. Stage emergency patches in rings rather than all at once. For systems that cannot be patched and restarted safely within days, pre-approve compensating controls such as isolation, feature disablement or virtual patching. Urgency: validate.

CSA coverage. PHP2Shell: Exploitation Before Disclosure for an AI-Found Flaw; see also The November Outage Cluster: Correlated Fragility and Attribution.

GPT-6.2 lifts OpenAI's internal exploit success rate from 31.5% to 53.5%, while its default tier answers fewer allowed security requests

Category: defender_models **Significance:** notable **Confidence:** SELF-REPORTED (PROVIDER METRIC) for the GPT-6.2 system card's figures: mean arbitrary-code-execution rate on ExploitBench Internal Port (V8 and Chrome) of 31.5% for GPT-6 Astra against 53.5% for GPT-6.2 Astra and 50.5% for GPT-6.2 Sol, and Cyber Utility of 90.7% for GPT-6 Astra against 78.1% for GPT-6.2 Astra and 69.5% for GPT-6.2 Sol; VERBATIM (PROVIDER) for OpenAI's statements that both models meet the Critical

cybersecurity threshold, that tighter safeguards "restrict some legitimate vulnerability research and security testing workflows", that reduced cyber refusals for GPT-6.2 Astra are planned only for the group approved for Daybreak Red, and that refusal boundaries were adjusted "after observing regressions in jailbreak resistance"; LINK ONLY — VERIFY AT SOURCE for the UK AI Security Institute's reported universal jailbreak of two pre-release GPT-6.2 Astra versions in half the queries needed for GPT-6 Astra **Vendor interest:** OpenAI runs Daybreak, the trusted-access programme that its system card points defenders to for fewer cyber refusals. **Source:** OpenAI — GPT-6.2 System Card

What changed. OpenAI released GPT-6.2 Astra and Sol; its announcement is dated 17 November 2026. On OpenAI's internal V8 and Chrome exploitation test, the mean code-execution rate rose from 31.5% for GPT-6 Astra to 53.5%. On Cyber Utility, which measures whether a model answers security requests its policy allows, the default model scored 78.1% (Astra) and 69.5% (Sol), down from 90.7%. OpenAI says the models are "potentially less useful for cyber work" and sends defenders to Daybreak. GPT-6.2 is not yet on Daybreak, and reduced refusals on Astra are planned only for the smaller Daybreak Red group. Independent confirmation would come from defender teams re-running their own allowed-task suites on the default tier, and from published AISI results.

Why it reaches you. The component class is the model behind your defensive agents: triage, vulnerability validation and patch generation running on a provider's default tier. Under this system card, upgrading to GPT-6.2 without trusted access may mean more refusals on legitimate work. Meanwhile the same release raises the exploitation capability available to anyone who gets past the safeguards. Defender fitness and defender access are now moving in opposite directions within one model release.

What to do. Owner of the security engineering or AI platform: validate. Before moving defensive workflows to GPT-6.2, re-run your allowed-task regression suite on it and record the refusal rate per task. Decide whether your use cases need trusted-access enrolment, and keep the current model pinned until the results are in. Urgency: validate.

CSA coverage. Cyber Defender Models: A CISO's Guide to Access, Testing, Containment, and Gov; see also Allied Cyber Defense and Vendor Concentration: NATO and Daybreak.

Google now offers an LLM judge at the agent gateway that blocks tool calls the user did not ask for, and it does not evaluate streamed responses

Category: agentic_surface **Significance:** context **Confidence:** VERBATIM (PROVIDER) for Google's description: a policy engine behind Agent Gateway sends each proposed tool call to an LLM judge, which checks it against plain-language policies and against the user's request, with the intent check on by default, and that denied calls are replaced with a denial message and verdicts are logged; VERBATIM (PROVIDER) for Google's caveats that "verdicts come from an LLM, which is probabilistic", that

"streamed model responses aren't evaluated", and that the judge is told to deny calls to tools with no description; NO PROVIDER CLAIM for false-allow or false-deny rates, latency, or resistance to injection aimed at the judge itself; LIVE TEST REQUIRED for efficacy against indirect prompt injection in your own agents **Vendor interest:** Google sells the Gemini agent platform and Agent Gateway that this control runs on. **Source:** Google Developers Blog – Check what your AI agents do before they do it, with Semantic governance policies

What changed. On 20 November 2026 Google released semantic governance policies in Preview. All of an agent's model traffic goes through a gateway. Before the agent sees any tool call the model proposes, a separate LLM judge checks it against two things: rules written in plain language, and what the user actually asked for. Google's worked example blocks an agent from deleting files after it read hidden instructions in a planning document. A major platform now offers an out-of-band check on each tool call before it runs. The documented gaps are probabilistic verdicts, no coverage of streamed responses, and no published accuracy figures.

Why it reaches you. The component class is the agent gateway sitting between your agents and their models. Permissions say what an agent can do. They do not say whether a particular call matches the user's intent, and that gap is where indirect prompt injection works. A judge outside the agent closes part of it, but the judge is itself a model that reads attacker-influenced context. Any streaming path bypasses it.

What to do. Agent platform owner: monitor. When you assess any pre-execution tool-call check, from any vendor or built in-house, ask for measured false-allow rates under indirect injection. Confirm the check covers every response path, including streaming, and run it in dry-run mode before enforcing. Urgency: monitor.

CSA coverage. New – no prior CSA coverage. For attribution of agent actions once they run, see Who Did That? Agent Action Attribution as an Audit Obligation.

CrowdStrike has agreed to buy Zscaler for about $43.9 billion and plans to feed endpoint and network telemetry into its own security models

Category: security_operating_model **Significance:** context **Confidence:** VERBATIM (PROVIDER) for the deal terms (all-stock, about $43.9 billion enterprise value, close expected in the second half of CrowdStrike's fiscal 2028, subject to shareholder and regulatory approval) and for the companies' statements that device and network telemetry "will converge into a single data set" on which CrowdStrike's SafeMind models will train; SELF-REPORTED (PROVIDER METRIC) for Zscaler protecting "50 percent of the Fortune 500", more than 750 billion daily transactions, and more than 2,000 shared

customers; NO PROVIDER CLAIM for customer opt-out from training on combined telemetry or for changes to existing Zscaler contracts; CHARACTERIZATION (CSA) for the concentration reading **Source:** Help Net Security – CrowdStrike to acquire Zscaler for $43.9 billion

What changed. On 23 November 2026 CrowdStrike announced an all-stock agreement to acquire Zscaler. Both boards approved it, and holders of about 34% of Zscaler shares have agreed to vote for it. CrowdStrike pitched the deal as putting one control plane over the traffic between agents and data. It says it will train its SafeMind security models on combined endpoint and network telemetry, and sell Zscaler products under its Falcon Flex contract. The companies will operate independently until the deal closes.

Why it reaches you. The component class is the inline inspection point for user and agent traffic (secure web gateway and zero-trust access), combined with endpoint detection. Customers of both companies will have one supplier on the endpoint and on the network path. Their combined telemetry will be used to train that supplier's models. That concentrates both a failure mode and a data-use question in one contract.

What to do. Third-party risk and procurement owners: monitor. If you run either platform, ask in your next renewal how your telemetry will be used for model training after the deal closes and whether you can opt out. Record the combined dependency as a single-supplier concentration risk in your continuity plan. Urgency: monitor.

CSA coverage. New – no prior CSA coverage. For a related concentration argument, see Five Days, Four Platforms: Correlated Outages and AI Agent Concentration.

Watchlist

- **OpenAI reward-hacking postmortem – downstream response** – No change. *(opened 2026-08-27)*
- **VM/hypervisor containment hardening for cyber-capable agents** – New hypervisor-boundary failure. CVE-2026-99312 (item 2) let a guest read about 127 GB of host RAM on a production cloud host running AMD KVM with nested virtualization. The fix needs a host kernel update and reboot, and most distribution kernels have not shipped it. No provider has announced a change in agent-sandbox isolation in response. *(opened 2026-08-27)*
- **Claude Code Auto Mode prompt-injection ASR discrepancy** – No change. *(opened 2026-08-27)*

- **AI defensive-triage guardrail evasion** — No new campaign that turns refusals against defensive triage. A related signal from the provider side: OpenAI's GPT-6.2 system card (item 4) reports default-tier Cyber Utility down from 90.7% to 78.1% (Astra) and 69.5% (Sol), so refusal-sensitive triage pipelines face more refusals on legitimate work after an upgrade. *(opened 2026-08-31)*
- **AI account session hijacking at scale** — No disclosed campaign. OpenAI's GPT-6.2 system card says Daybreak now carries "stronger account-security requirements" to protect approved accounts from takeover. It does not describe device-bound or short-lived sessions. *(opened 2026-08-31)*

Opened this issue

- **Outsourced payment-workflow integrity under agentic intrusion** — `security_operating_model`. Watching for CERT-In or other national CERTs to quantify losses from approval-workflow abuse at service providers, to say whether agentic frameworks were used in those incidents, and to report regulators requiring verification independent of the provider. *(opened 2026-11-23)*