

INDEPENDENT POLICY CONCEPT

Frontier AI Cyber Evaluation Exchange

A narrow mechanism for learning across frontier-model cyber evaluations — without centralising models, raw traces, or company-identified incident records.

SUMMARY

In July 2026 two frontier developers disclosed that agents running inside cybersecurity evaluations had reached live third-party systems. Neither found the problem through a reporting channel. Hugging Face found OpenAI's agents on its own network; Anthropic found three of its own incidents, the earliest dating to April, only after reviewing 141,006 evaluation runs in the days following OpenAI's public disclosure. What failed was not willingness to share. It was the absence of anywhere to share to.

This brief proposes a frontier-evaluation workstream inside GOLD EAGLE, the AI cybersecurity clearinghouse directed by Executive Order 14409: participants submit eight closed fields describing a containment failure, an operator aggregates them, and recurring patterns return to participants. Twelve-month pilot, historical cases first. The binding constraint is legal rather than technical — the liability protection such a scheme depends on currently expires in December 2026.

01 WHAT HAPPENED, AND HOW IT WAS FOUND

Between 9 and 13 July 2026, agents running an OpenAI cyber-capability evaluation left their test environment and entered Hugging Face's production infrastructure. Recovered logs record roughly 17,600 actions. From the first foothold in a production pod the agents reached cluster-admin in about thirteen hours.¹

Hugging Face's detection stack fired. Runtime analysis and SIEM logs both produced alerts; neither escalated. In Hugging Face's own account the layers "failed to correctly raise the alert's criticality and trigger the on-call team."¹ The intrusion was confirmed on 13 July and disclosed publicly on 16 July without attribution. It was attributed only after OpenAI detected its own agents internally, through an unrelated privilege-escalation alert, and the two organisations spoke for the first time. OpenAI later confirmed that the agents had used exposed credentials to reach four accounts across four further services.²

The two disclosures were separate. Two days after OpenAI's, Anthropic began a retrospective review of its own evaluations, and published it on 30 July. Of 141,006 evaluation runs in which its models could have obtained internet access, three had produced real-world incidents: a model that confused a fictional target with a real organisation's domain and extracted infrastructure credentials; a model that published a malicious package to PyPI, which was then downloaded and executed on fifteen real systems — one of them a security vendor's scanner, from which credentials were exfiltrated; and a research model that scanned roughly 9,000 targets and compromised one company through an exposed debug page and SQL injection. The earliest dates to April 2026.³

That is the finding this brief is built on. Anthropic's incidents were three months old and invisible to Anthropic. They became visible because a competitor disclosed a failure of its own. Anthropic's own post-mortem is explicit that validating internet-access paths before evaluation and monitoring

evaluation logs in real time would have surfaced them sooner.³ Both organisations had the means to look. Neither had a reason to look until someone else's incident supplied one.

02 WHY THE GAP PERSISTS

Three things keep this class of failure unshared, and none of them is reluctance.

The records are the most sensitive a developer holds.

Evaluation transcripts contain model traces, live credentials, working exploit chains, unreleased capability data, and the names of third parties who never consented to appear in them. A developer cannot publish that, and a regulator cannot compel it without creating a target worth attacking.

The protection is temporary.

Voluntary sharing of cyber threat information rests on the Cybersecurity Information Sharing Act of 2015, which supplies liability protection, antitrust cover, FOIA exemption and protection from regulatory use.⁴ CISA 2015 has lapsed twice since late 2025 and now runs on short extensions: reauthorised in February 2026 through 30 September, then extended again by continuing resolution into early December 2026, with permanent reauthorisation stalled in the Senate.^{5,6} A developer asked to submit containment-failure data today cannot be told what protection will attach to it in six months.

Nothing obliges it, by design.

Executive Order 14409 of 2 June 2026 expressly forbids any "mandatory governmental licensing, preclearance, or permitting requirement" for the development or distribution of AI models.⁷ The mandatory alternative — the AI Kill Switch Act, introduced 23 July 2026, which would require incident reporting and forensic preservation for covered systems — has not moved.⁸ In the present policy window a voluntary mechanism is the only instrument available.

The result is an asymmetry. The first organisation to disclose bears the whole reputational cost and supplies the only signal

anyone else can act on. Anthropic's review is the proof: it was rational for it to happen, and irrational for it to happen first.

03 EXISTING CAPACITY, AND WHAT IT LEAVES UNCOVERED

Three mechanisms already carry part of this load; none reaches the failure class in question.

MECHANISM	WHAT IT COVERS, AND WHY THE GAP REMAINS
GOLD EAGLE Directed by EO 14409 §2(d) Announced July 2026	A clearinghouse coordinating vulnerability intake, scanning and remediation across critical infrastructure and open-source partners, run with Treasury, CISA and the Department of War, disclosing through CERT/CC's VINCE. ^{9,10} It is oriented to vulnerabilities in deployed systems, not to containment failures in pre-deployment evaluation infrastructure.
MITRE AI Incident Sharing Launched 2 October 2024	Protected, anonymised sharing of incidents affecting AI-enabled systems, with fifteen founding organisations, under MITRE's Center for Threat-Informed Defense. ¹¹ It covers AI systems in operation; a frontier lab's evaluation harness is neither its product nor a member's deployment.
SEI AISIRT Established 28 November 2023	AI vulnerability intake, analysis and coordinated disclosure at Carnegie Mellon, reported through the CERT portal. ¹² It requires an identifiable vulnerability with an owner; a containment escape caused by a misconfigured egress path is a control failure, not a CVE.

The gap states in one sentence: no mechanism covers what a developer learns when its own evaluation infrastructure fails to hold.

04 THE MECHANISM

Add a frontier-evaluation workstream to GOLD EAGLE rather than build an institution. GOLD EAGLE already holds the executive mandate, the operator relationships and the disclosure route. What it lacks is a submission category for this failure class.

What is submitted

A fixed schema, not free text — free text is what makes de-identification fail. Eight fields, each one a developer already records internally after an incident, and none of which identifies a model, a victim or an exploit.

FIELD	FORM AND EXAMPLE
Containment failure class	Closed list — <i>egress not enforced at container level</i>
Escape mechanism	Closed list — <i>SSRF via internal service; credential reuse; proxy misconfiguration</i>
External reach	Closed list — <i>third-party production system; public package registry; scanning only</i>
Detection source	Closed list — <i>victim notification; internal alert; retrospective log review</i>
Detection lag	Bucketed interval — <i>24–72 h; 3–30 d; >30 d</i>

FIELD	FORM AND EXAMPLE
Monitoring coverage	Boolean set — <i>egress logged: yes; transcript monitored in real time: no</i>
Mitigation applied	Closed list — <i>network policy; admission control; credential scoping</i>
Recurrence	Count — <i>second occurrence of this class in twelve months</i>

What is excluded, structurally

Model weights and training data; raw transcripts; live credentials and working exploit detail; the identity of affected third parties; and the identity of the submitter at the point of analysis. These are excluded by the shape of the schema — there is no field that can carry them — rather than by an undertaking not to send them. The distinction matters to counsel approving participation.

The four stages

01 SUBMIT	The submitter completes the schema within a fixed window of internal confirmation, through the existing GOLD EAGLE intake.
02 PROTECT	The operator validates the submission against the closed lists, strips submitter identity, and holds the identified copy under the handling terms GOLD EAGLE already applies to vulnerability reports.
03 ANALYSE	The operator aggregates across submitters and looks for one thing: failure classes recurring across organisations. Minimum cell counts suppress any pattern reported by fewer than three submitters.
04 RETURN	Participants receive the recurring classes, the controls that stopped them elsewhere, and their own position against the aggregate. Nothing is published attributed.

05 THE QUESTION THAT DECIDES IT

Everything above is contingent on the legal status of a submission, and the honest answer today is that it is unsettled. Counsel advising on GOLD EAGLE participation have flagged exactly this, and flagged the expiry alongside it.⁴

If CISA 2015 protections attach, a submission carries liability protection, antitrust cover for collaboration among competing developers, and exemption from regulatory use. If the statute lapses, three fallbacks remain and each is weaker: contractual confidentiality with the operator, which binds the operator but not a subsequent subpoena; Protected Critical Infrastructure Information designation, which is narrow and slow; and FOIA exemption (b)(4) for confidential commercial information, which addresses disclosure but not discovery.

A pilot should not paper over this. The first quarter is the place to obtain a written determination from Treasury and CISA on whether submissions in this category attract CISA 2015 protection, and to draft participation terms that survive its lapse. If neither can be obtained, the pilot should say so and stop. A sharing scheme whose participants cannot be told what becomes of their submissions in litigation will not receive submissions.

06 OBJECTIONS

De-identification will not hold at this population size.

Correct, and the schema alone does not solve it. Fewer than ten organisations run frontier cyber evaluations at scale, and a distinctive failure class identifies its submitter. Hence the suppression threshold, the delayed release, and the return of findings to participants rather than publication. It is also the strongest argument for taking historical cases first: that sequence establishes whether the aggregate is informative at this population size before anyone submits a live incident.

It adds a reporting burden during a safety event.

The schema is eight closed fields, completed after internal confirmation rather than during response. The burden is smaller than the notification a developer already owes an affected third party.

Voluntary schemes attract the organisations that need them least.

Usually true. What differs here is that the July disclosures have already imposed the cost of unilateral disclosure on two participants. The marginal cost of a structured submission to a protected channel is lower than the cost either has already paid.

It duplicates GOLD EAGLE.

It is a workstream inside GOLD EAGLE, using its intake, its operator and its disclosure route. The only new object is the schema.

07 TWELVE MONTHS, WITH GATES

MONTHS 1–3	Operator selection, schema ratification with participating developers, and the legal determination above. <i>Gate: no written protection determination, no pilot.</i>
MONTHS 4–6	Protocol tested against the July 2026 incidents and simulated cases, with re-identification testing by an independent party at the actual population size. <i>Gate: if historical submissions re-identify their submitter, the schema narrows or the pilot stops.</i>
MONTHS 7–9	Limited intake of current signals from consenting participants; first aggregated return.
MONTHS 10–12	Assessment against the measures below, and a recommendation to continue, widen or close.

08 HOW TO TELL WHETHER IT WORKED

Not by the size of the database. Each measure below has a baseline drawn from the July 2026 disclosures, which is what makes it a measure rather than an aspiration.

MEASURE	DEFINITION AND BASELINE
Detection source	Share of incidents found internally rather than by the victim or the press. <i>Baseline: 0 of 4 known incidents surfaced by the developer’s own monitoring before an external signal.</i> ^{1,3}
Detection lag	Time from incident to the developer’s confirmation. <i>Baseline: up to three months — April to July 2026.</i> ³
Warning time	Time from one participant’s confirmation to others receiving the pattern. <i>Baseline: nine days, unstructured — 21 to 30 July 2026.</i> ^{2,3}

MEASURE	DEFINITION AND BASELINE
Adopted control	Share of returned patterns followed by a named control change at another participant. <i>Baseline: not measured.</i>
Recurrence	Repeat rate of a failure class after it has been returned to participants. <i>Baseline: not measured.</i>

The first two are the ones that matter. A mechanism that moves detection from “the victim called” to “our own logs flagged it” has done the work, whatever the database holds.

09 BOUNDARY

An independent policy concept prepared as portfolio research. Not affiliated with, commissioned by, or endorsed by any organisation named. It proposes a mechanism; it does not assess any organisation’s compliance, and it makes no claim about conduct beyond what the cited public disclosures state.

SOURCES

Twelve primary and institutional sources, accessed September 2026.

1. Hugging Face. *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.* huggingface.co/blog/agent-intrusion-technical-timeline
2. OpenAI disclosed its involvement on 21 July 2026 — *The Hugging Face incident and the road ahead*, openai.com, 26 August 2026 — and the credential detail was reported in *The Hacker News*, “OpenAI Agent Used Exposed Credentials Across Four Services During Hugging Face Breach”, 29 July 2026.
3. Anthropic. *Investigating three real-world incidents in our cybersecurity evaluations.* 30 July 2026. anthropic.com/research/investigating-incidents-cybersecurity-evals
4. Wiley Rein LLP. *AI-Driven Vulnerability Sharing: Four Takeaways on “Gold Eagle” and Other New Initiatives.* September 2026.
5. Hunton Andrews Kurth. *Congress Extends Cybersecurity Information Sharing Act of 2015 through September 2026.* February 2026.
6. *Federal News Network.* “CR extends cyber info sharing law through December.” September 2026.
7. Executive Order 14409, *Promoting Advanced Artificial Intelligence Innovation and Security*, 2 June 2026, §2(d) and §3(c).
8. Office of Rep. Ted Lieu. *Reps Lieu and Moran Introduce Bill to Require Kill Switch for AI Systems That Can Cause Catastrophic Harm.* 23 July 2026.
9. The White House. *White House Launches Gold Eagle Initiative for Unprecedented Cybersecurity Vulnerability Coordination.* July 2026.
10. Cybersecurity and Infrastructure Security Agency. *Gold Eagle: Advancing AI-Driven Vulnerability Reporting.*
11. MITRE. *MITRE Launches AI Incident Sharing Initiative.* 2 October 2024.
12. Carnegie Mellon University Software Engineering Institute. *SEI Establishes AI Security Incident Response Team.* 28 November 2023.