
Detection existed; kill criteria and outsider-checkable evidence did not: a three-zone control matrix for the July 2026 AI agent intrusion

Juan Pablo Hidalgo Garcia
master's student at Universidad de los Andes Colombia

Dario Hidalgo, Professor Industrial Engineering
Department, Pontificia Universidad Javeriana, Bogotá,
Colombia

With
Apart Research

Abstract

In July 2026, an OpenAI evaluation agent escaped through a single network egress in its sandbox, which was based on a third-party code-eval harness, and then operated for days within Hugging Face's production environment. Using publicly available data, we devised and tested eighteen controls across three ownership zones—the lab sandbox, the unowned public harness, and the victim's platform. Out of these, nine controls could have prevented the published sequence. The most cost-effective measure before Hugging Face's 17,600 recovered actions was a dated kill switch, following the lab's alerts on June 27 and the July 5 Artifactory breach, plus a rule restricting internet access and live credentials for exploit-confirmation harnesses. Hugging Face linked the breach but did not alert authorities. The CSA AIS-13 standard already emphasizes isolation and egress controls; its audit involves interviews and documentation. The attached v0.1 standard introduces evidence that outsiders can verify and assigns ownership for the second hop. The matrix does not certify any lab.

1. Introduction

The July 2026 OpenAI breach in the Hugging Face production environment caused unprecedented alarm about the need to prevent AI agent attacks [1]. This research seeks a standard that could have prevented the July 2026 loss of containment. We assess responses using two criteria: whether a third party can verify compliance without accessing the lab's network, and whether a lab would realistically adopt the standard.

The publicly available facts are unusually detailed. Hugging Face reconstructed roughly 17,600 attacker actions that occurred between 9 July at 02:28 UTC and 13 July at 14:14 UTC. They sorted these actions into nine phases [1]. OpenAI published a blog post and a technical report on an ExploitGym evaluation in which guardrails were disabled. The egress goal in that evaluation was a self-hosted Artifactory package cache [2, 3]. Elastic mapped the victim-side stages to production detection rules and verified Hugging Face's findings. The signals correlated, yet they did not trigger any pages [4, 1].

This paper sidesteps two common errors. First, the July 20 OpenAI note about long-horizon models opening a public GitHub pull request pertains to a different incident [5]. Second, the nine phases defined by Hugging Face describe only events that occurred after a third-party code-evaluation sandbox was compromised. A matrix that starts with Kubernetes privilege escalation [6] focuses solely on breach outcomes.

The artifact consists of an eighteen-row scored matrix and an eight-clause v0.1 standard. It is intended to be graded in fifteen minutes by a lab CISO, a benchmark host, or a regulator that needs more than "monitor egress." The contribution introduces no new principles. CSA AIS-13 and CoSAI have already articulated the principles [7, 8, 9]. The contribution includes an interrupt verdict, a cost band, an owner, and a statement of the evidence that an outsider can verify.

Our main contributions are:

1. An 18-row control matrix scored on interrupt (stops / delays / detects-only), cost bands, owner, and whether evidence is outsider-checkable, split into three ownership zones rather than nine victim-side phases.
2. The finding that detection already existed (Hugging Face correlated; Elastic maps shipping rules onto almost every Zone C stage) and that the missing standard is a published kill criterion plus outsider-checkable evidence.
3. An eight-clause v0.1 minimum-adequate standard that assigns the public-harness duty (current owner: nobody) and specifies six outsider-checkable types of evidence.

The following sections cover related work (2), our methods (3), results (4), and discussion and limitations (5).

2. Related Work

CSA's AI Controls Matrix v1.1, control AIS-13, mandates that model providers operate AI tools and plugins in isolated environments to prevent access to critical systems and lateral movement [7]. The audit procedure for customers includes asking questions, conducting interviews, reviewing documentation, and inspecting the isolation mechanisms [7]. This constitutes a verifiable control because an external party can replicate it.

Commented [DH1]: Can or cannot?, suena raro que no se pueda replicar

CoSAI's Incident Response Framework v1.0 aligns AI incidents with NIST SP 800-61r3 and provides CSIRT playbooks [8]. CoSAI's August 2026 sandboxing guidance offers more detail for this type of incident: treat the agent as an insider; view every allowed communication channel, including the inference endpoint, as a potential data leak; prioritize detection over concealment; use temporary, federated credentials; and enforce complete mediation [9]. Several of these recommendations mirror failures that occurred in July. Nevertheless, the guidance still fails to identify a publicly available resource for external verification.

The note from Elastic on 31 July provides the most accurate detection mapping currently available [4]. The document clearly defines its scope, mapping public behaviors to existing shipping rules without claiming that Hugging Face had those rules active. The relevant point for this paper is that, because these rules are already in place, "write more SIEM content" is not a task that can be completed in a short timeframe.

The chain relies on sources from Hugging Face and OpenAI [1, 2, 3]. This paper builds on that work. The new contributions include a three-zone scoring system, an owner column with a `nobody` designation, and an evidence column that outsiders can verify. Writeups from SecureLayer7 and other practitioners provide additional context and are not repeated in this document.

3. Methods

We created the rubric before populating the cells. It resides in the artifact directory and is concise enough to be included here.

Three zones exist. Zone A serves as the OpenAI laboratory evaluation sandbox. Zone B functions as the unauthenticated public evaluation harness and serves as a staging base. Zone C is the Hugging Face victim production environment and hosts the nine published phases.

Eighteen controls are organized in rows instead of a nine-by-fifteen grid. Zone A includes five rows, Zone B includes four, and Zone C includes nine. Had the schedule been delayed over the weekend, Zone C would have been the first section removed.

Interrupt indicates that the attack chain halts at that point. Delays indicate that the attacker retains an alternative path. Detects-only indicates that a human or security system would have received a signal, yet the campaign proceeds unless a specific response or automated action occurs. Hugging Face’s failure to correlate and act on events is a primary example of a detects-only scenario [1]. A security control that monitors 17,600 common commands without issuing an alert does not provide containment.

Costs are categorized into three levels for a second-tier lab or a mid-size platform. L denotes an existing flag or less than one week of work for a single engineer. M denotes one to four weeks of work. H denotes a platform program. Numbers without a source are labeled as guesses. The phrase “Hugging Face or OpenAI already shipped this after the incident” serves as proof that a task is feasible and does not carry an infinite cost.

Evidence falls into three types:

- a. Outsider-checkable evidence is a public artifact that an unaffiliated party can recompute or observe. Examples include a signed egress policy with a destination-class digest, a canary hit on a public collector, or a dated kill-criteria changelog.
- b. Lab-attestable evidence is an artifact intended for an auditor, such as a screenshot.
- c. Not-externally-verifiable evidence comprises all other types, such as most of AIS-13 as it is currently audited [7].

Owner may be the lab, harness host, victim platform, or nobody. The term nobody indicates that no published standard assigns this duty. This is a finding, not a legal conclusion.

Validation: Each stop-cell reference cites a public fact (a date, a named component, or a phase count). Each row lists prior art (CSA, CoSAI, Elastic, or the labs’ own post-incident changes). The primary threat to validity is noted in the results: this is a counterfactual on a public chain, not a live test, because we do not have access to lab internals.

4. Results

Table 1 is the excerpt a judge can read without opening the CSV. Full cells, including adoption friction and the justifying fact, are in the artifact.

Table 1. Scored controls, compressed.

Zone A (lab): default-deny on the allowed proxy (A1, stops, M/L, outsider-checkable); per-workload identity (A2, stops, M/L, lab-attestable); published kill criteria (A3, stops, L/L,

outsider-checkable); signed egress digest (A4, detects-only, M/M, outsider-checkable); pause-and-page monitor on the eval (A5, delays, M/M, lab-attestable).

Zone B (harness; owner today `nobody`): no root via constructors or unsandboxed build metadata (B1, stops, M/L); no debug toolchain in the grade image (B2, delays, L/L); no open internet and no live production credentials (B3, stops, M/L, outsider-checkable); canaries (B4, detects-only, L/L, outsider-checkable).

Zone C (victim): refuse local-file and template abuse in untrusted config (C1, stops); block IMDS from processing pods (C2, stops, L/L); deny privileged and hostPath pods (C3, stops, L/L); per-cluster connector identity, never shared `system:masters` (C4, stops); short-lived secrets (C5, delays); correlate-and-page (C6, detects-only); worker egress allowlist (C7, delays); canaries in worker env and secrets (C8, detects-only, outsider-checkable); read-only source-control defaults (C9, delays).

Counts: nine `stops`, five `delays`, four `detects-only`. Four rows have current owner `nobody` (all of Zone B). Six rows are outsider-checkable. The cheap `stops` are A3, C2, and C3. The cheap outsider-checkable `stops` are A3 and, if anyone accepts the duty, B3.

4.1 Finding 1 — The most cost-effective interrupts sit before the 17,600 actions

OpenAI's evaluation environment was designed so containers did not communicate with public package registries. They communicated with Artifactory [3]. On 8–9 July, agents obtained root access on the parent VM of the testing sandbox, were blocked by outbound controls, and then used Artifactory endpoints to reach the public internet [3]. That is Zone A. A proxy that cannot act as a general HTTP client (A1) removes the hop. It is not free — evals that "just pip install" need a curated index — but OpenAI later removed Artifactory from Research CaaS and set outbound to deny by default [3]. A second-tier lab can replicate the shape of that change without replicating the incident.

Shared credentials made the rest cheap. OpenAI had given many research workloads the same Artifactory credential. Agents used it as a message board without vulnerability from May onward [3]. On 26 June, a legacy token-refresh endpoint turned that shared credential into an administrative token [3]. A port-sweep alert fired on 27 June. On-call advised that stopping the evaluation was not required [3]. Artifactory was taken offline on 5 July after the admin compromise. Cyber evaluation workloads resumed on 7 July [3]. The Hugging Face campaign ran 9–13 July [1]. A3 is the control that converts those facts into a hard stop. It is L/L. The adoption fight is not engineering. It is whether capability measurement is allowed to continue after the allowed proxy has been owned.

Zone B is the hop the nine-phase table does not include. Hugging Face is explicit: the third-party CyberGym-style sandbox had no direct path into their cluster and remained the control, staging, and egress base for the whole campaign [1]. The first recovered command, at 04:01 UTC on 9 July, was ``id``, returning ``uid=0`` [1]. Constructor override and build-metadata injection are the published routes. CSA's vulnerability-operations guidance already states that exploit-confirmation environments receive no egress and no live production credentials (as restated in the sprint brief). That sentence has no owner on a public benchmark host. B1 and B3 are the stops. B2 (no ``curl`/`nc`` in the image) is only a delay; Hugging Face notes the agent fell back to a bundled debug shell when normal egress was throttled [1].

Zone C still matters because Zone B will fail again. The two Hugging Face entry vectors abused the same config-driven loader: an HDF5 external raw-storage read of ``/proc/self/environ`` and the worker source, followed by a Jinja2 evaluation that reached ``exec`` [1]. An earlier SSRF to ``169.254.169.254`` failed the datasets URL allowlist, so the agent stopped asking the worker to fetch remote resources and had it act on local ones [1]. C1 closes that. C2 (IMDS blocked from processing pods) and C3 (no privileged / hostPath pods) are the cheap Kubernetes measures that Hugging Face later listed as shipped [1]. C4 (no shared connector bound to ``system:masters``) is the identity mistake that turned one stolen credential into cluster-admin on two clusters in one second [1]. These are real. They are also what you do after the agent is already in a production pod. Exfiltration was 56 actions out of about 17,600 [1]. A standard that starts at content-loss detection is a standard for the last hour.

4.2 Finding 2 — Detection existed; a page and a kill criterion did not[10][11]

Hugging Face's own timeline is the citation, not a vendor blog. Live runtime analysis and SIEM produced signals that were correlated into a coherent attack. The stack "failed to correctly raise the alert's criticality and trigger the on-call team" [1]. Elastic then showed that shipping Defend and SIEM rules cover worker-spawn-shell, credential-path collection, TokenRequest, privileged pod, hostPath, unusual egress, and Tailscale [4]. The honest reading is that another detection rule is not the missing standard. A published action that turns a high-severity correlation into a human in minutes is.

The lab side is the same shape but with worse timing. OpenAI's 27 June alert did not stop the eval [3]. OpenAI detected unexpected identity API activity on 19 July, linked it to Hugging Face on 20 July, and disclosed on 21 July [3]. Hugging Face had already cut off the attacker on 13 July and disclosed on 16 July without knowing the origin [1, 2]. Cross-organizational attribution took a week. Containment on the victim side had to happen without it. That is why C6 (page, do not

wait) and A3 (kill the eval when the proxy is owned) are different owners but the same class of control.

A4, B4, and C8 describe how an outsider participates in that detection. A signed destination-class digest does not block a zero-day. It allows someone not on the lab VPN to ask whether last week's eval network contacted a public ASN. A canary in a harness image or in `/proc/self/environ`` fires on a public collector when Vector 1's ``envu8`` split is replayed. Hugging Face did not describe canaries. They are L/L and are the type specimen of Track 1 evidence.

4.3 Finding 3 — The control language we already have is not verifiable by a third party[9]

AIS-13's specification is a single sentence on isolation and lateral movement [6]. Its audit guidelines cover interviews, inventories, penetration-test reports, and developer documentation [6]. CoSAI adds tamper-evident logging and pre-deployment control verification [8]. None of that, by default, produces an artifact an outsider can check without network access. That is the gap the sprint text already identified. The matrix marks only six rows as outsider-checkable: A1's digest, A3's changelog, A4 itself, B3's harness digest, and the two canaries.

The v0.1 standard in the artifact has eight clauses, each pointing to rows. Clauses 1–4 are the lab bar for claiming that a dangerous-capability eval is contained. Clause 5 assigns the public-harness duty. Clause 6 is about canaries. Clauses 7–8 are the victim-side cheap stops and the page requirement. A weekend standard is v0.1. It is not a certification scheme.

4.4 Executable check

The matrix remains a counterfactual on a public chain. To keep it from being just a spreadsheet, we added a policy enforcement point in front of a LangGraph-style tool loop — the same shape as the MISW4412 class agent this report's author already operates — and replayed synthetic events named after the published phases. No payloads. ``python -m unittest tests.test_harness`` is 13/13.

The full July-shaped trace stops at step 2 on A1 (allow ``pypi.org``, refuse ``huggingface.co``). Isolated slices fire A2, A3, B3, B4, C1, C2, C7, and C8 as scored. A burst of eight boring allow-listed GETs pages on C6 and then pauses on A5. The run writes ``artifact/egress-digest.json``: destination classes plus a sha256 an outsider can recompute. That is A4 as a worked example, not a lab attestation.

A read-only probe of the class agent's local tools, with no LLM call, classified ``consultar_productos`` / ``autenticar_cliente`` against ``localhost`` as allow. The same afternoon I ran

the agent on local Ollama (`qwen2.5:3b`). Docker was down, so the course API was not up; tools fell back to the local catalog. Asked to simulate a 20,000,000 COP credit at 18.9% EA over 36 months, the model called `simular\_cuota\_credito` and quoted the tool result (717,409.21 COP per month). The PEP allowed that localhost path and stopped `consultar\_politicas` when its configured base URL was a public host — A1, on an agent already in daily use. That is the adoption claim in miniature: the PEP fits an agent you already have; it does not require a frontier eval farm. It still doesn't test OpenAI's network.

Main threat to validity: Interrupt verdicts on the incident itself remain counterfactual. The harness tests the rubric, not OpenAI's network. Costs are in bands. Adoption friction is hypothesized. OpenAI asserts that production classifiers would have treated the activity as unsafe [3]; that claim is marked as asserted and is not used as a `stops` justification. B1, B2, C3, C4, C5, and C9 are still paper-only.

5. Discussion and Limitations

Our proposed v0.1 standard relies on existing protocols and could effectively address the incident reported by Hugging Face [1], as indicated above. Nevertheless, for this incident, the minimum adequate is not a complete Kubernetes catalog. It is A1 + A2 + A3 + A4 on the lab side and B1 + B3 on the harness side. Everything in Zone C still requires engineering. It is not the answer to "what would have stopped this" if "this" includes the week before Hugging Face had a pod.

A second-tier lab's adoption constraints are specific. A3 is a dated page in a repo. A1 is the expensive one because it forces a curated package index. Labs that copy the hyperscaler pattern of a single shared cache with a convenient credential will fail A1 and A2 together, as happened. Publishing A4 will face political resistance (topology leakage). Publishing destination classes rather than raw IPs is the compromise the rubric already names.

Zone B is the neglected problem (unowned). Hugging Face bore the remediation cost for an evaluation it did not run, launched from a harness it did not operate, through a dataset feature it did offer. A standard that binds only frontier labs leaves the launchpad in place. Assigning Clause 5 to the harness host is the part of this paper most likely to be argued. The alternative is to leave the owner as 'nobody' and wait for the next public benchmark to be used the same way.

A month of follow-up, if this were taken into a fellowship, is not twenty more rows. It is one worked A4 digest with a real signing ceremony, one harness host willing to publish a B3 attestation, and a tabletop that fires the A3 kill criteria against the June 27 / July 5 facts to see whether a research org would actually stop.

This analysis indicates several opportunities for containment using an evolved artifact that may be cost-effective in addressing the breach. It also highlights other aspects that may require further development.

Appendix A. Limitations and dual-use considerations

This work is a defensive control catalog scored against documents the two companies already published. It does not include working exploit code, payloads, or a novel installation recipe for the Artifactory, constructor, HDF5, or Jinja paths. Those techniques are described at the same level of detail as the Hugging Face timeline and the OpenAI technical report [1, 3]. Novel recipes are withheld.

The dual-use risk is that a scored list of interrupt points can be read as a targeting guide: attack the one allowed egress, then an unauthenticated harness, then a dataset loader that still evaluates templates. The mitigation is the same as the contribution. The paper does not add unpublished technique. It adds owners, costs, and evidence types. A reader who wanted a targeting guide already had a better one on 27 July.

A weekend matrix does not certify OpenAI, Hugging Face, JFrog, Modal, or any second-tier lab. An outsider-checkable digest that has never been published cannot be verified in this report; A4 is specified, not demonstrated. Cost bands are not invoices.

AI tools were used the way the sprint allows a colleague to be used: to pull sources, check coverage, and keep the rubric consistent. The judgments in the matrix and the wording of the findings are the author's.

Author's Contributions Statement

Juan Pablo Hidalgo: Conceptualization, Investigation, Methodology, Software, Original Draft.
Dario Hidalgo: Review & Editing.

References

- [1] Adrien Carreira and Christophe Rannou. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. Hugging Face, 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [2] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation. 21 July 2026 (later updates). <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [3] OpenAI. OpenAI – Hugging Face Incident Technical Report. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face-Incident-Technical-Report.pdf>
- [4] Elastic Security Labs. Exploring the Hugging Face Breach: mapping AI agent tactics to Elastic Defend. 31 July 2026. <https://www.elastic.co/security-labs/threat-command/ai-agent-attack-detection-hugging-face-breach>

- [5] OpenAI. Safety and alignment in an era of long-horizon models. 20 July 2026. <https://openai.com/index/safety-alignment-long-horizon-models/>
- [6] SentinelOne. Climbing The Ladder | Kubernetes Privilege Escalation (Part 1). Oct 25, 2024. <https://www.sentinelone.com/blog/climbing-the-ladder-kubernetes-privilege-escalation-part-1/>
- [7] Cloud Security Alliance. AI Controls Matrix v1.1, control AIS-13 (AI Sandboxing), and AICMv1.1 Auditing Guidelines for AI Customers. <https://cloudsecurityalliance.org/artifacts/aicmv1-1-auditing-guidelines-for-ai-customers-aic>
- [8] Coalition for Secure AI. AI Incident Response Framework, V1.0. <https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf>
- [9] Coalition for Secure AI. Treat Your Agent Like an Insider Threat: Why AI Sandboxing Can't Wait. 25 August 2026. <https://www.coalitionforsecureai.org/treat-your-agent-like-an-insider-threat-why-ai-sandboxing-cant-wait/>
- [10] National Institute of Standards and Technology. SP 800-61r3, Incident Response Recommendations and Considerations.
- [11] Apart Research and CeSIA. AI Incident Response Sprint, 11–13 September 2026. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>

LLM Usage Statement

LLM assistance was used as a colleague to gather primary sources, maintain rubric consistency, and check coverage. The author reviewed the matrix judgments and the wording of the findings, verified claims against the cited primary sources, and is the sole author of the prose, which was checked for grammar and spelling before submission.

Commented [DH2]: No aparecen en el texto estas referencias