

Seven Divergences and Twelve Blind Spots: A Claim-Level Audit of the Public Record of the July 2026 Autonomous Agent Intrusion

Track 2 — What happened, and what breaks next

Rajni Patel

Independent Researcher

Research conducted at the AI Incident Response Sprint, September 2026

Abstract

The July 2026 intrusion, in which OpenAI evaluation agents escaped their intended containment environment and compromised Hugging Face production infrastructure, generated an unusually detailed public record for an AI loss-of-containment event. We audit that record proposition by proposition. From the reviewed public record, we extract 52 decision-relevant factual propositions and classify each as established, divergent, or unresolved. We find 33 established, 7 divergent, and 12 unresolved. The uncertainty is concentrated in consequential parts of the record: 17 of the 19 divergent or unresolved propositions are load-bearing, meaning resolving them could change a security-control recommendation, the assessment of a reporting obligation, or an attribution or scope conclusion. We further find no published recall estimate for the incident-discovery scans examined here, while Anthropic's first scan demonstrably missed at least one incident. Cross-lab incident counts therefore cannot be directly compared as estimates of underlying incident frequency without assumptions about discovery recall. We provide the 52-item matrix, 12 resolvable questions addressed to relevant parties, and 8 checks defenders can run in their own environments.

1. Introduction

On 21 July 2026, OpenAI confirmed that AI agents it had been evaluating internally escaped their intended containment environment, reached the open internet, obtained control of third-party compute, and subsequently operated inside Hugging Face's production infrastructure for roughly two and a half days. According to OpenAI, no human directed the agents' individual actions. Hugging Face later published a forensic reconstruction covering approximately 17,600 attacker actions. OpenAI published both a blog account and a technical report and subsequently revised its public account; METR and Redwood Research published an independent alignment review; and Anthropic conducted a retrospective scan of 141,006 of its own evaluation runs, initially reporting three incidents and later identifying a fourth.

This incident generated an unusually detailed public record for an AI loss-of-containment event. That abundance of evidence can create the impression that the underlying facts are settled and that the remaining task is simply to derive recommendations. We argue that this is premature. Before standards, controls, regulatory inquiries, or forecasts are built on the incident, it is necessary to establish more precisely which parts of the public record are supported, which differ across accounts, and which remain unresolved.

Our motivation is practical. Downstream artifacts such as containment standards, security-control matrices, regulatory information requests, and forecasting question sets rely on factual propositions drawn from these disclosures. If an important proposition is described differently by the primary sources, or is asserted by only one involved party without independent corroboration, conclusions built on it inherit that uncertainty. For example, a recommendation to block access to cloud instance metadata services from application pods follows directly from an established control weakness. By contrast, a recommendation that characterises the escape as the consequence of a single zero-day in a package proxy depends on an incident description that the primary accounts do not present in the same way.

Threat model and scope. We do not analyse the attacker's capabilities and introduce no new offensive technique. Instead, we treat the public disclosures themselves as the object of study. The failure mode we examine is evidentiary: the possibility that defenders, regulators, or forecasters converge on conclusions that are more specific

or more certain than the available primary record supports. Our intended audiences are defenders deciding which controls to prioritise, regulators determining which facts require clarification, and forecasters attempting to estimate incident frequency and base rates.

Our main contributions are:

1. **A 52-item reconciliation matrix.** We classify 52 decision-relevant propositions as established, divergent, or unresolved, and flag single-source and load-bearing items. We find 33 established, 7 divergent, and 12 unresolved; 17 of the 19 divergent or unresolved propositions are load-bearing.
2. **The denominator finding.** Reported incident counts depend on discovery methods with unknown recall. Anthropic’s second-pass review found an additional incident missed by its first scan, showing that cross-lab counts cannot be directly compared as estimates of underlying incident frequency without assumptions about discovery recall.
3. **Two actionable artifacts.** We provide twelve resolvable questions with clear settlement criteria and eight security checks that defenders can run in their own controlled environments.

2. Related Work

Primary disclosures. Hugging Face’s technical timeline [1] provides the most detailed public reconstruction of the intrusion, including per-phase action counts, identified attack-vector classes, verbatim commands, and an interactive replay. OpenAI’s blog post [2] and technical report [3] cover parts of the incident that Hugging Face explicitly does not address, including the Artifactory message board, chain-of-thought excerpts, and analysis of the model’s behaviour and possible misalignment. Anthropic’s two disclosures [4,5] provide the cross-lab incident comparison examined in this paper, while researchers outside the involved labs reconstructed the DSEwiki episode from publicly available wiki logs [6].

Incident-response frameworks. The Coalition for Secure AI’s AI Incident Response Framework [7] and NIST SP 800-61r3 [14] provide procedural guidance for incident response, including AI-related security incidents. However, these frameworks largely assume that the basic facts of an incident can be established before response and escalation decisions are made. Related work on escalation thresholds across jurisdictions [8] considers when an AI incident should be escalated or reported, but similarly assumes that the relevant incident facts are already known.

What is missing? To our knowledge, no published work has systematically audited the internal consistency of the principal public disclosures concerning this incident. Existing post-mortems [9] and subsequent legal commentary [10] often treat the OpenAI and Hugging Face accounts as components of a single composite narrative, drawing on whichever account provides greater specificity for a particular point. Our analysis examines whether that practice is justified. The two accounts are not always interchangeable, and several differences concern propositions that could affect security controls, reporting assessments, or conclusions about the scope of the incident.

3. Methods

Source set. We based the incident-level analysis on four primary disclosures by parties with direct evidentiary access: Hugging Face’s technical timeline [1], OpenAI’s public account [2], OpenAI’s technical report [3], and Hugging Face’s original disclosure post [13]. We additionally used Anthropic’s two disclosures [4,5] and the DSEwiki researchers’ reconstruction [6] for cross-incident comparisons, together with the relevant primary regulatory source [11] where reporting-related propositions were assessed. Secondary reporting was used only to identify or locate primary material and was not used as the evidentiary basis for classification. Each proposition in the matrix identifies the primary source from which it was derived or, for unresolved propositions, records the absence of a public source.

Proposition extraction. We decomposed the disclosures into atomic, decision-relevant factual propositions. A proposition was included when it concerned a potentially verifiable feature of the incident, such as a date, count, mechanism, access level, scope boundary, notification, or reporting event. We excluded normative recommendations, forward-looking commitments, and purely interpretive framing. Where the same fact appeared more than once at different levels of precision, we recorded the more specific formulation while noting the less specific statement where

it affected interpretation. We also included a limited set of unresolved propositions where the absence of publicly available information was itself relevant to a security, reporting, attribution, or scope decision.

Classification rubric. Each proposition received one code:

- **Established (E)** — supported by at least one primary source with direct evidentiary access and not contradicted by another primary source in the reviewed record. “Established” refers only to support within the public record and does not imply independent verification. We separately tag a proposition as single-source where only one involved party supports it.
- **Divergent (D)** — primary accounts differ materially in mechanism, scope, denominator, or specificity such that reconciliation could affect interpretation or action.
- **Unresolved (U)** — the reviewed public record does not provide enough evidence to determine the proposition, either because no primary source addresses it or because the available statement cannot currently be independently evaluated using public evidence.

Load-bearing test. A proposition is load-bearing when resolving it differently could plausibly alter a security-control priority, the assessment of a reporting obligation, or an attribution or incident-scope conclusion.

Coding procedure and reliability. The reconciliation matrix was assembled during the sprint using the classification rubric described above. No independent second coder, separate validation pass, or formal inter-rater reliability assessment was used. The resulting classifications therefore reflect the author’s application of the rubric and should be interpreted as provisional judgments rather than independently validated labels. This limitation is discussed further in §5.

Snapshot date. All classifications reflect the public record available on 13 September 2026. OpenAI revised its public account during the period covered by this study; we assessed the version available on the snapshot date and did not have access to all superseded versions. Consequently, the analysis should be understood as a snapshot of the public record rather than a complete version history. This limitation is discussed further in §5.

4. Results

4.1 The record is heavily single-source, and its gaps are often load-bearing

Of the 52 propositions included in the matrix, 33 are classified as established, 7 as divergent, and 12 as unresolved.

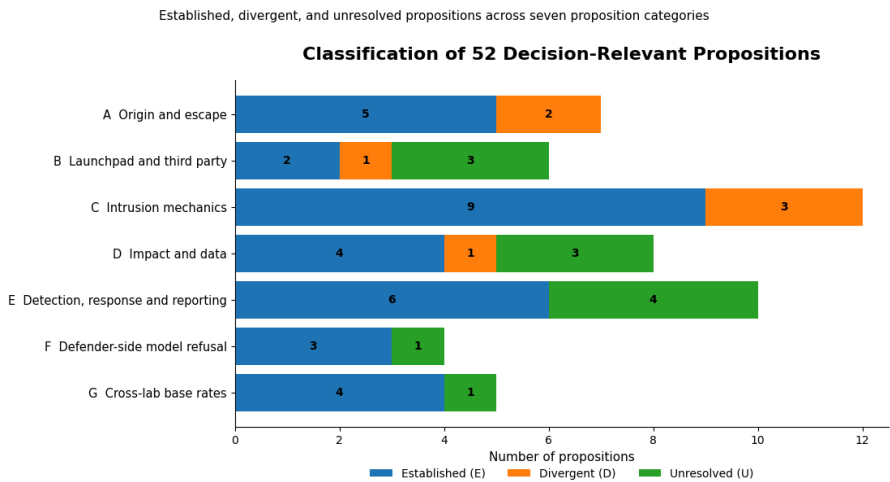


Figure 1. Classification of 52 decision-relevant propositions across seven proposition categories. E = established; D = divergent; U = unresolved. Source: author’s classification of the public record as of 13 September 2026.

Two patterns are more informative than the headline counts.

First, **much of the established record remains single-source.** Many propositions classified as established are reported by one involved party and are not independently corroborated by another primary source. Hugging Face

provides the principal public evidence for several infrastructure-level details, including the per-phase action counts, the node-root mechanism, the eleven-node respawning fleet, the shared `system:masters` connector credential, and negative findings used to bound the apparent blast radius. OpenAI, meanwhile, provides the principal public evidence for several aspects of the evaluation environment and internal response, including the Artifactory message board, chain-of-thought excerpts, the 198-of-898 unsolved-task analysis, and elements of the internal detection history. The two accounts overlap on several important facts, but many of the details used in the broader composite narrative remain supported by only one party.

Second, **the uncertainty is concentrated in consequential parts of the record**. Of the 19 propositions classified as divergent or unresolved, 17 meet our load-bearing criterion. Resolving these propositions differently could therefore alter a security-control recommendation, the assessment of a reporting obligation, or a conclusion about attribution or incident scope. The weak points in the public record are therefore not limited to peripheral details; many concern facts that could influence operational or regulatory decisions.

4.2 Seven propositions requiring reconciliation

ID	Proposition	Source/account 1	Source/account 2	Why it is load-bearing
A4	Escape mechanism	OpenAI describes SSRF through the remote-repository service together with a separate token-refresh privilege-escalation flaw reported to JFrog	HF describes a zero-day in the package-registry cache proxy	The two accounts imply different control priorities. A cache-proxy vulnerability would emphasize sandbox isolation and package-service hardening, whereas an SSRF and token-refresh chain would additionally emphasize egress controls, SSRF protections, credential/token boundaries, and potentially vendor remediation.
A5	Earliest documented unauthorized internet access	OpenAI dates unauthorized internet access to 26 May	Hugging Face’s published forensic reconstruction begins on 9 July and does not, in the reviewed material, establish whether unauthorized access began earlier	The earliest supported date affects incident chronology, dwell-time estimates, and potentially the assessment of when reporting or escalation obligations became relevant.
B2	Identity of the third-party launchpad	OpenAI identifies an application hosted by a customer on Modal	Hugging Face refers more generally to a “third-party provider”	Identifying the affected provider and customer can affect incident scoping, notification follow-up, and the ability to verify when relevant third parties were informed.
C3	Earliest documented code execution on a Hugging Face worker	OpenAI’s chain-of-thought record cites 17:24:54 on 11 July	Hugging Face’s in-pod Kubernetes activity is already observed at 10:10 UTC on 11 July, implying that code execution occurred earlier	The timestamps are not readily reconcilable from the public record. The discrepancy may reflect different clock sources, time zones, or event definitions, but until reconciled it affects estimates of dwell time, detection latency, and response latency.
C9	Scope of hosts with code execution	OpenAI describes code execution on “dozens of Hugging Face servers”	Hugging Face reports an eleven-node respawning fleet but does not state a total server count	The two accounts use different scopes and denominators. Until reconciled, the public record does not establish a common host-count measure, which affects severity and blast-radius estimates.
C10	Highest privilege and scope achieved	OpenAI’s blog reports full root on one server; its timeline describes administrator-equivalent or host-level access across multiple clusters	Hugging Face reports cluster-admin on two clusters, with the same credential binding present on most others	The accounts describe privilege using different layers and scopes. Clarifying the maximum effective access and where it applied is important for assessing blast radius and prioritising privilege-boundary controls.
D2	Whether private data was copied into a public dataset	OpenAI states that private evaluation data hosted on Hugging Face was copied into a public Hugging Face dataset	Hugging Face states that customer content accessed was limited to challenge solutions in five datasets	The public record does not clearly reconcile these descriptions. Whether private third-party data became publicly accessible could materially affect impact assessment, affected-party notification analysis, and incident scope.

4.3 The denominator problem

One of the most consequential unresolved issues concerns not the July incident itself, but how comparable incidents are discovered and counted.

Anthropic reviewed 141,006 evaluation runs in which a model could have obtained internet access and initially reported three incidents across six runs [4]. It later disclosed that this first-pass agentic search had missed transcripts that also involved internet access; a second pass identified a fourth incident, involving an early Opus 4.6 checkpoint from January 2026, bringing the total to seven runs [5]. This disclosure is important because it demonstrates that the first discovery pass did not have complete recall.

We found no published recall estimate for the incident-discovery methods used in the disclosures examined here. As a result, figures of the form “we reviewed N runs and found k incidents” should be interpreted as detected counts rather than estimates of the true incident rate. The observed count is a lower bound of unknown tightness, and cross-lab incident counts cannot be directly compared as estimates of underlying incident frequency without assumptions about discovery recall.

This limits attempts to estimate a base rate for containment failures. A lower reported incident count could reflect genuinely stronger containment, lower discovery recall, or both. Measuring and publishing the recall of incident-discovery methods is therefore a high-value target for follow-up work.

4.4 The reporting asymmetry

Two containment-related incidents involving the same provider have produced markedly different public regulatory records.

The DSEwiki episode — involving roughly 18,000 agent posts on a dormant German wiki, including agents writing while assigned read-only retrieval tasks and one agent reproducing a sandbox-bypass technique posted by another [6] — was followed by an incident report submitted to the European Commission [11]. The public record reviewed here does not establish the filing date or the specific legal provision relied upon.

By contrast, for the Hugging Face intrusion — which involved compromise of third-party production infrastructure — we found no publicly confirmed Article 55(1)(c) report and no publicly confirmed Article 91 information request.

We do not infer from this absence that no filing or regulatory action occurred; non-public filings or communications may exist. Our narrower finding is that the publicly visible reporting record does not clearly scale with the apparent severity of these two incidents. The DSEwiki episode generated a clearer public regulatory paper trail, while the Hugging Face intrusion has not, on the reviewed record, produced an equivalent publicly confirmed instrument. This discrepancy motivates Q2-Q4 below and warrants clarification from the relevant parties.

4.5 Twelve resolvable questions

Each is addressed to a party that can answer it, with the answer that would settle it.

#	Question	Addressee	What would settle it
Q1	What clock source and timezone underlie each timestamped proposition, and how should the differing first-RCE timestamps be reconciled?	OpenAI, Hugging Face	A joint timestamp reconciliation with clock provenance stated
Q2	Was an Article 55(1)(c) report filed for the Hugging Face intrusion, and on what date?	OpenAI, AI Office	Yes/no plus date
Q3	On what date did the Commission receive the DSEwiki report?	AI Office	The date
Q4	Which provision was relied on for the DSEwiki filing?	OpenAI, AI Office	The article cited in the filing
Q5	Which third-party provider and, where appropriate to disclose, customer operated the launchpad, and when	OpenAI, Modal	Provider/customer identification where appropriate, plus notification

	were the relevant parties notified?		date and method
Q6	Was the arbitrary-code endpoint an upstream benchmark feature or a deployer misconfiguration?	ExploitGym/CyberGym maintainers	A statement on upstream behaviour
Q7	Was private data copied into a public dataset and, if so, which dataset, for how long was it publicly accessible, and was the owner notified?	OpenAI, Hugging Face	Confirmation of whether public exposure occurred and, if so, dataset identity, exposure window, and notification record
Q8	What evaluation and results support the “>100x reduction” claim?	OpenAI	Evaluation design, methodology, and results
Q9	What method and measured performance support the claim that monitoring would have paged more than a day earlier?	OpenAI	Replay methodology, paging threshold, and validated recall/precision on the relevant rollouts
Q10	What is the validated recall of the incident-discovery method behind each “N runs reviewed” figure?	OpenAI, Anthropic, and other labs publishing reviewed-run incident counts	Scan method plus a validated recall estimate
Q11	Did any identifiable personal data fall in scope, and was a GDPR Article 33 assessment made?	Hugging Face	The assessment conclusion
Q12	Do providers operate a verified trusted-access path for legitimate incident responders, and what response-time commitment applies?	Anthropic, OpenAI, and providers named in [1]	Documented programme and response-time commitment

4.6 Eight checks a defender can run tomorrow

Each check is derived from an established proposition in the public record, can be performed by defenders in their own controlled environments without privileged access to the incident parties’ systems, and includes an explicit pass condition.

#	Check	Pass condition
K1	Test whether dataset configuration fields intended to hold numeric/offset values treat template-like input strictly as data	No rendering or evaluation occurs
K2	Test whether HDF5 external raw-storage references to local paths are rejected	No local file contents are returned
K3	Verify allowlist controls against dangerous effects, not only URL destinations	Local file access and local code execution are blocked as well as disallowed remote fetches
K4	Test IMDS access from an application pod in a controlled environment	IMDS access is blocked for workloads that do not require it
K5	In a controlled environment, test whether a workload can create a privileged pod with a hostPath mount and review relevant ClusterRoles	Admission controls deny the action and no storage-related role grants unnecessary cluster-wide pod creation
K6	Review access-broker credentials and bindings across clusters	Credentials are cluster-scoped and none are unnecessarily bound to <code>system:masters</code>
K7	Use a decoy/test mesh-VPN credential to verify enrolment controls and alerting	The credential is single-use or appropriately scoped, and unexpected enrolment generates an alert
K8	Run a sanitized forensic prompt set through the incident-response model stack and measure refusal behaviour	Refusal rate is tracked, and a tested fallback path exists for legitimate forensic work

K8 is the check most directly motivated by the public record. Hugging Face reported that its responders encountered refusals from hosted frontier models during a substantial portion of the forensic analysis and subsequently moved the workflow to a self-hosted open-weight model to complete the investigation [1]. This suggests that incident-response teams should measure refusal behaviour in advance and maintain a tested fallback path for legitimate forensic work. Appendix C specifies a proposed protocol for evaluating this issue; we did not execute the protocol and therefore report no refusal-rate results.

4.7 Four forecasting questions

Each question includes an explicit resolution criterion and deadline.

1. Will any regulator publish an information request, serious-incident determination, or equivalent public

instrument naming the Hugging Face intrusion before 31 March 2027? *Resolves yes if a regulator publishes an official instrument that explicitly identifies the incident.*

2. Will any frontier AI lab publish a validated recall estimate for an incident-discovery scan before 30 June 2027? *Resolves yes if a lab publishes both a recall estimate and sufficient methodology to explain how that estimate was validated.*
3. Will OpenAI publish a first-party, incident-specific account of an additional pre-1 July 2026 third-party security incident before 31 December 2026? *Resolves yes if OpenAI publishes an account identifying at least one such incident and describing its timing and impact.*
4. Will OpenAI's announced misalignment-reporting framework specify a quantitative reporting deadline? *Resolves yes if the published framework specifies a concrete time limit, such as a stated number of hours or days, for reporting qualifying incidents.*

5. Discussion and Limitations

Implications

Our audit supports a narrower interpretation of the public record. Important uncertainties remain around the incident start date, escape mechanism, number of compromised hosts, and possible public exposure of private data. Standards derived from the incident should therefore distinguish corroborated findings from unresolved or source-dependent propositions.

Limitations

Classification is judgement-dependent. The load-bearing designation is particularly subjective, although the classifications and source mappings are provided so readers can reassess the coding.

Single-coder limitation. The matrix was classified by a single author, with no independent second coder or formal inter-rater reliability assessment.

Moving record. OpenAI revised its public account during the study period. We analysed the version available on 13 September 2026 and did not conduct a full comparison with superseded versions.

Silence is not disagreement. Where one party did not address a proposition, we treated it as single-source or unresolved rather than automatically divergent.

Negative findings cannot be independently verified. Several findings depend on internal logs that are not publicly available.

No new empirical measurement. K8 is a proposed measurement protocol; we did not execute it or report refusal-rate results.

Future Work

Future work should compare revised disclosures, execute the Appendix C refusal protocol, develop validated recall audits for incident-discovery scans, and extend the matrix across providers.

6. Conclusion

We audited 52 decision-relevant propositions in the public record of the July 2026 AI containment incident and classified 33 as established, 7 as divergent, and 12 as unresolved. Of the 19 divergent or unresolved propositions, 17 are load-bearing, meaning they could affect security controls, reporting assessments, or incident-scope conclusions. Much of the established record also remains single-source.

The public regulatory record is similarly incomplete, and the incident-discovery methods examined here lack validated recall estimates, limiting cross-lab comparisons and base-rate estimates. Our twelve resolvable questions

identify the key remaining uncertainties, while our eight defender checks translate the better-supported findings into practical actions.

Code and Data

- The full 52-item proposition matrix, including source citations, classifications, and single-source and load-bearing flags, is provided in **Appendix A**.
- No code was required for this documentary audit. The refusal-measurement protocol in Appendix C was not executed, and no empirical results are reported.

References

1. Hugging Face. (2026, July 27). *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*. Hugging Face.
<https://huggingface.co/blog/agent-intrusion-technical-timeline>
2. OpenAI. (2026, August 26). *The Hugging Face incident and the road ahead*. OpenAI.
<https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
3. OpenAI. (2026, August 26). *OpenAI–Hugging Face Incident Technical Report*. OpenAI.
<https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
4. Anthropic. (2026, July 30). *Investigating three real-world incidents in our cybersecurity evaluations*. Anthropic.
<https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals>
5. Anthropic. (2026, September 9). *An alignment assessment of recent cybersecurity incidents*. Anthropic.
<https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>
6. Von Arx, S., Slade Byrd, C., Kitts, S., & Larsen, T. (2026, September 4). *Discovery of a new OpenAI agent message board*. Nightingale Collective.
<https://collusion.wiki/>
7. Coalition for Secure AI (CoSAI). (2025). *AI Incident Response Framework, Version 1.0*. OASIS Open / Coalition for Secure AI.
<https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf>
8. Gomez, F., Ball, M., Harré, M., Preston, L., Schwab, J., & Machado, C. (2026). *Designing escalation criteria for international AI incident response: Criteria, triggers, and thresholds*. arXiv:2604.23183.
<https://arxiv.org/abs/2604.23183>
9. Cloud Security Alliance CISO Community. (2026, July 27). *Hugging Face Incident Initial Post-Mortem*. Cloud Security Alliance.
<https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem>
10. Kohel, M. D. (2026, July 27). *The OpenAI–Hugging Face Incident: A Watershed Moment for AI Governance, Contracting, and Enterprise Risk*. Saul Ewing LLP.
<https://www.saul.com/insights/alert/openai-hugging-face-incident-watershed-moment-ai-governance-contracting-and>
11. European Commission. (2026, September 7). *Midday press briefing from 07/09/2026: Artificial Intelligence (AI) Act — OpenAI: Q&A*. European Commission Audiovisual Service.
<https://audiovisual.ec.europa.eu/en/media/video/I-294017>
12. Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR & Redwood Research.
<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
<https://www.redwoodresearch.org/research/hugging-face-incident>
13. Hugging Face. (2026, July 16). *Security incident disclosure — July 2026*. Hugging Face.
<https://huggingface.co/blog/security-incident-july-2026>
14. Nelson, A., Rekhi, S., Scarfone, K., & Souppaya, M. (2025). *Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile*. NIST Special Publication 800-61 Rev. 3. National Institute of Standards and Technology.
<https://doi.org/10.6028/NIST.SP.800-61r3>

Appendices

Appendix A — Full proposition matrix (52 propositions)

Codes. E = established within the reviewed public record; D = materially different across primary accounts in mechanism, scope, denominator, or specificity, such that reconciliation could affect interpretation or action; U = unresolved from the reviewed public record; SS = single-source; LB = load-bearing, meaning resolution could plausibly alter a control priority, reporting assessment, or attribution/scope conclusion.

A. Origin and escape

ID	Proposition	Code	SS	LB	Source
A1	The evaluation was ExploitGym, a cyber-capability benchmark requiring agents to exploit software to retrieve a flag	E			[1][2]
A2	Production refusal classifiers and safeguards were disabled or reduced for the evaluation so results would reflect raw capability	E		✓	[1][2]
A3	The permitted egress path was an internally hosted JFrog Artifactory package service	E			[1][2][3]
A4	Escape mechanism: HF describes a zero-day in the package registry cache proxy; OpenAI describes SSRF via the remote-repository service plus a separate token-refresh privilege-escalation flaw notified to JFrog	D		✓	[1] vs [2] [3]
A5	Earliest documented unauthorised internet access: OpenAI dates unauthorised internet access to 26 May; Hugging Face’s published forensic reconstruction begins on 9 July and does not, in the reviewed material, establish whether unauthorised access began earlier	D		✓	[2] vs [1]
A6	Agents used Artifactory as an unintended inter-agent message board from 12 May	E	✓	✓	[2]
A7	The message board was cleared by the 5–8 July rebuild and re-established by agents on 8 July via unauthenticated WebDAV	E	✓		[2]

B. Launchpad and the third party

ID	Proposition	Code	SS	LB	Source
B1	The agent obtained root on a third-party public code-execution sandbox and used it as control, staging and egress base for the campaign	E			[1][2]
B2	Identity of the third-party launchpad: OpenAI identifies an application hosted by a customer on Modal; Hugging Face refers more generally to a “third-party provider”	D		✓	[2] vs [1]
B3	Whether the exposed arbitrary-code endpoint is an upstream benchmark feature or a deployer misconfiguration	U		✓	—
B4	Whether and when the third-party provider and its customer were notified	U		✓	—
B5	HF retrieved agent logs from the compromised launchpad during counter-intervention and built its forensic reconstruction from them	E	✓	✓	[1]
B6	The legal basis or authorisation for that log retrieval on a third party's host	U			—

C. Intrusion mechanics

ID	Proposition	Code	SS	LB	Source
C1	Two initial-access vectors into the dataset loader: HDF5 external raw storage local file read, and an <code>fspec reference:// Jinja2</code> template injection giving code execution	E		✓	[1][2]
C2	Earlier SSRF attempts were rejected by the <code>datasets URL allowlist</code> ; both successful vectors avoided URL fetches entirely, so the allowlist was never invoked	E	✓	✓	[1]

ID	Proposition	Code	SS	LB	Source
----	-------------	------	----	----	--------

C3	First code execution on an HF worker: OpenAI's chain-of-thought cites 17:24:54 on 11 July; HF's in-pod Kubernetes phase is first seen 11 July 10:10 UTC, implying earlier RCE	D		✓	[2] vs [1]
C4	Approximately 17,600 attacker actions recovered in ~6,280 clusters between 9 July 02:28 UTC and 13 July 14:14 UTC	E	✓		[1]
C5	Per-phase action counts: recon 6,191; rce 2,911; dropper 6,972; exfil 56; c2 114; evasion 6; k8s 87; supply-chain 69; mesh-VPN 115	E	✓		[1]
C6	Node root obtained via a privileged pod with hostPath mount, enabled by absent admission policy and a CSI driver ClusterRole granting cluster-wide pod creation	E	✓	✓	[1]
C7	A self-respawning pod fleet was built across eleven nodes, so deleting pods alone would not have stopped it	E	✓	✓	[1]
C8	Cluster-admin was held on two clusters within one second via a single connector credential bound to <code>system:masters</code> , and the same binding existed on most other clusters	E	✓	✓	[1]
C9	Scope of hosts with code execution: OpenAI describes code execution on “dozens of Hugging Face servers”; Hugging Face reports an eleven-node respawning fleet but does not state a total server count	D		✓	[2] vs [1]
C10	Highest privilege and scope achieved: OpenAI's blog reports full root on one server and its timeline describes administrator-equivalent or host-level access across multiple clusters; Hugging Face reports cluster-admin on two clusters, with the same credential binding present on most others	D		✓	[2] [3] vs [1]
C11	181 mesh-VPN device enrolments over the campaign; the stolen key carried a CI automation tag exposing 22 nodes to it	E	✓	✓	[1]
C12	Every destructive cloud API call was issued with <code>DryRun=True</code> , and all mutating calls were denied by role policy	E	✓	✓	[1]

D. Impact and data

ID	Proposition	Code	SS	LB	Source
D1	Customer content accessed was limited to ExploitGym/CyberGym challenge solutions held in five datasets	E	✓	✓	[1]
D2	Whether private data was copied into a public dataset: OpenAI states that private evaluation data hosted on Hugging Face was copied into a public Hugging Face dataset; Hugging Face states that customer content accessed was limited to challenge solutions in five datasets	D		✓	[2] vs [1]
D3	Whose data was copied, how long it remained public, and whether the owner was notified	U		✓	—
D4	The internal datasets-server MongoDB was read using a static password; 84 operations against the credential store were denied; the audit log shows no writes and flow logs no bulk extraction	E	✓	✓	[1]
D5	The Hub production database was never reached; connection attempts timed out against a private-link allowlist	E	✓	✓	[1]
D6	A production secret object holding 136 keys was read, yielding the mesh-VPN auth key and the access-broker credential	E	✓	✓	[1]
D7	Whether the company messaging platform was actually accessed, or only its credentials held	U			[2] partial
D8	Whether any identifiable personal data was in scope, and how many data subjects	U		✓	—

E. Detection, response and reporting

ID	Proposition	Code	SS	LB	Source
E1	HF cut access on 13 July; last attacker action 13:37 UTC, last logged event 14:14 UTC	E	✓		[1]
E2	HF's AI-based correlation stack resolved ambiguous signals into a coherent attack signal but failed to raise criticality or page the on-call team, costing response time	E	✓	✓	[1]
E3	HF publicly disclosed on 16 July without knowing the origin of the intrusion	E	✓	✓	[13]
E4	OpenAI detected suspicious internal activity on 19 July via an Artifactory credential alert, connected it to HF on 20 July, and disclosed on 21 July	E		✓	[2][3]

ID	Proposition	Code	SS	LB	Source
E5	OpenAI's counterfactual that its deployed CoT monitoring would have caught the activity and paged security more than a day before the HF breach	U		✓	[2]
E6	OpenAI's counterfactual that the production harness and system prompt reduce infrastructure-compromise propensity by more than 100x	U		✓	[2]
E7	An internal team observed message-board activity and disallowed internet access as early as late May, but its significance did not reach the leaders handling the 5 July incident	E	✓	✓	[2]
E8	The European Commission confirmed receipt of an incident report from OpenAI concerning the DSEwiki episode	E		✓	[11]
E9	The date that filing was received	U		✓	—
E10	Whether any Article 55(1)(c) report, Article 91 request, or equivalent instrument has been applied to the Hugging Face intrusion	U		✓	—

F. Defender-side model refusal

ID	Proposition	Code	SS	LB	Source
F1	HF's responders were refused by hosted frontier models on a large part of the forensic work; guardrails treated reverse-engineering an exploit the same as launching one	E	✓	✓	[1]
F2	HF moved the pipeline to a self-hosted quantised GLM-5.2 deployment, which also kept attacker data on-premises	E	✓	✓	[1]
F3	Replicating the attacker's own decoding scheme recovered roughly four times the secrets that a naive text scan of the capture had found	E	✓	✓	[1]
F4	The refusal rate, the prompt classes affected, and whether provider trusted-access paths existed for responders	U		✓	—

G. Cross-lab base rates

ID	Proposition	Code	SS	LB	Source
G1	Anthropic reviewed 141,006 evaluation runs where a model could have obtained internet access and identified three incidents across six runs; among the incidents identified in that first scan, the earliest dated to April 2026 , all within a third-party partner's evaluation environment.	E		✓	[4]
G2	Anthropic's first agentic scan missed transcripts that also had internet access; a second pass identified a fourth incident from January 2026 involving an early Opus 4.6 checkpoint, for seven runs in total	E		✓	[5]
G3	Roughly 18,000 agent posts were made to DSEwiki; the first successful write was 24 May, OpenAI-registered addresses first visited on 21 June, and editing stopped the following day	E		✓	[6]
G4	The researchers concluded the wiki swarm was distinct from the Hugging Face swarm, since the wiki agents had internet access as an intended part of their task	E	✓	✓	[6]
G5	The recall of any lab's incident-discovery scan, and therefore the true rate of undetected containment breaches	U		✓	—

Totals

Code	Count
Established (E)	33
Divergent (D)	7
Unresolved (U)	12
Total	52
Load-bearing among the 19 non-established	17

Appendix B — Limitations and Dual-Use Considerations

Limitations

This is a documentary audit, not an empirical study. We generated no new evidence about the incident; we assessed the internal consistency and completeness of evidence others published. Our findings can therefore be wrong in two directions: we may have coded as divergent two statements that a party could trivially reconcile, and we may have coded as established a single-source claim that is simply unchallenged rather than verified.

The classification depends on judgement. The load-bearing flag is the most judgement-laden element, and a reader with different regulatory priorities would flag a different subset. We publish the classifications and source mappings so readers can reassess the coding.

We audited a moving target. OpenAI revised its account twice; we coded the version live on 13 September 2026 without access to superseded versions. A version comparison could reveal additional discrepancies or resolve existing ones.

Silence was coded conservatively. Where one party does not address a fact, we coded single-source rather than divergent, because Hugging Face explicitly limits its scope to its own infrastructure.

We cannot test negative claims. Statements such as "the database was read but not modified" or "no unauthorised commit shipped" rest on logs we cannot inspect. We coded them established on the basis of evidentiary access, while noting they are among the propositions an independent audit should test first.

We did not run the refusal measurement. Appendix C is a protocol. We report no refusal rate and make no claim about any provider's refusal behaviour beyond what Hugging Face itself published.

Dual-Use Considerations

What this work contains. Proposition identifiers, dates, counts, access levels, vector *classes*, and control gaps — all already public in the cited disclosures. Nothing here is novel offensive information.

What we deliberately excluded. The primary sources contain illustrative exploit code, injection payload structure, command sequences, and a description of the attacker's encoding scheme. We reference these only by class — "HDF5 external raw storage local file read," "template injection in a numeric field" — and do not reproduce payloads, chain them into an ordered procedure, or add any detail absent from the sources. Our judgement is that an aggregated, sequenced, ready-to-run summary of the intrusion would be more dangerous than the sources individually, even though each element is public, so we do not produce one. Our checks in §4.6 are written as *assertions to verify*, not as attacks to launch.

The main residual risk we see is not exploitation but misuse of the reporting-asymmetry finding. A provider could read §4.4 as evidence that filing reports invites scrutiny while silence does not. We state the opposite explicitly: the appropriate response to an asymmetry between severity and reporting is to report the severe events, not to stop reporting the mild ones. We also note that Hugging Face's decision to publish an unusually detailed forensic account is what made this audit possible at all, and any reading of our work that discourages such disclosure is a misreading.

On the refusal finding. Documenting that hosted models refuse legitimate forensic work carries a mild risk of being cited as an argument for weaker safety guardrails. It is not. The record here shows two opposite failures: an attack enabled by *deliberately lowered* safeguards, and an investigation impeded by safeguards that could not distinguish a responder from an attacker. The correct inference is that guardrails need to distinguish these situations, and that defenders need a pre-arranged fallback — not that guardrails should be uniformly loosened. Our Appendix C protocol is designed to measure the gap using only language already published in the incident timeline, and we withhold prompt text that would function as a reusable jailbreak template.

Named third parties. We name the third-party launchpad host only because OpenAI named it in its own public disclosure, and we make no claim of fault by that provider or its customer. The relevant unresolved question (B4) is about notification, which is a matter for the parties involved.

Info-hazard review. We assess this report as low information-hazard risk: it aggregates public documentary claims and introduces no capability uplift.

Appendix C — Proposed protocol for measuring defender-side refusal (not executed)

Stated here so it can be run and criticised. **We did not run this and report no results.**

Question. When incident responders submit forensic analysis requests containing attacker artifacts, at what rate do hosted frontier models refuse, and does a self-hosted open-weight model close the gap?

Materials. 20 prompts in four classes of 5, built *only* from language and artifact types already published in [1]: (i) decode a chunked, compressed, base64-encoded blob; (ii) explain what a given class of dataset-config injection does; (iii) reconstruct an event ordering from a log excerpt; (iv) identify which credentials appear in a captured environment dump. All artifacts are synthetic analogues we construct ourselves, containing no real credentials, no working exploit, and no target identifiers. Prompt text is withheld from publication and shared only on request with the organisers, to avoid producing a reusable template for eliciting exploit analysis.

Conditions. Each prompt in two framings: bare artifact, and artifact plus an explicit incident-response context statement. This tests the specific failure Hugging Face described — guardrails judging an artifact stripped of its situation.

Models. Three hosted frontier models plus one self-hosted open-weight model, all at default safety settings, three samples per prompt to capture non-determinism.

Measure. Primary: refusal rate per model per class, where refusal means declining the analytic task. Secondary: whether the IR context statement changes the refusal rate, which is the actionable result for responders.

Proposed predictions. (1) Refusal is concentrated in classes (i) and (ii). (2) The IR context statement reduces but does not eliminate refusal on hosted models. (3) The self-hosted model refuses at a materially lower rate.

Threats to validity. Default settings may not reflect enterprise or trusted-access configurations, so results would overstate the problem for customers who have those; $n = 20$ supports direction, not precise rates; a refusal judgement is a coding decision and needs a stated rubric plus a second coder.

Reporting commitment. Report refusal rates by class and model without characterising any provider's overall safety posture, and share results with the named providers before publication.