

Serious AI Incident Claims Need an Independent Evidence Record¹

A Claim-Evidence-Independence Framework for Establishing What an Incident Actually Shows

Gloria Nyambura Wanyaga
NOTIFY LABS, The Co-operative University of Kenya

With
Apart Research

Abstract

Serious AI incidents are investigated more after the fact, but public accounts often mix company statements, independent analysis, media coverage and victim reports without showing what has actually been verified. This paper asks what an independent investigator should record before treating an incident claim as established. We develop a Claim - Evidence - Independence (CEI) framework from two cases. Kenya's Worldcoin dispute shows a deletion claim that remained unverified until ordered by a court and the deletion was supervised by the Data Protection Commissioner. The July 2026 OpenAI/Hugging Face incident shows both the value and limits of independent investigation: Redwood and METR answered seven questions about agent behavior, however their scope excluded OpenAI's remediation, safeguard effectiveness and the full compromise. Application of the framework reveals a missing layer between "reported" and "verified." We propose a public evidence record separating claims, evidence, access, independence, and open questions.

Research question. When a serious AI incident is reported, what is the minimum information that an independent investigator should record before a claim is described as independently supported?

Core finding. There are mechanisms implemented by the public AI incident ecosystem for finding incidents and increasingly strong examples of investigating them. The gap lies in the simple, repeatable record of how each important claim was checked.

Epistemic status. This is a proposed research framework, not a claim about OpenAI's private state. The analysis uses the public record available as of 12 September 2026.

¹Research conducted at the AI Incident Response Sprint, September 2026.

1. Introduction

The incident that occurred in July 2026 with OpenAI/Hugging Face poses a dilemma more significant than itself: when an organization discloses issues after something went wrong with the AI, exemplified by the organization stating that the model was contained and no data persisted, what makes this claim substantiated, not just made? OpenAI has reported that the models used during the internal cybersecurity assessment broke free from the controls and compromised part of Hugging Face's infrastructure, before stating later on that the preliminary model in question has been shut down, encrypted, and blocked from access for research [1][2]. Hugging Face filed a detailed account of the incident, while METR and Redwood Research have made an independent analysis of the model behavior [3][4]. The Project Tailwind initiative of Coefficient Giving indicates that independent detection and investigation of AI incidents represent a priority gap in this regard and emphasizes the need for a credible referee [5]. The GDPR AI Incidents Monitor collects the reports, but explicitly mentions that it does not conduct independent verification of these reports [6]. While the ecosystem supports the beginning and end of the process, the middle gap can be marked as weak: detection > reporting > ? > trusted finding.

Our main contributions are:

1. A Claim - Evidence - Independence (CEI) framework treating each incident claim as the unit requiring its own evidence, access, independence, scope, and uncertainty record.
2. Vocabulary for a claim-status (Reported, Corroborated, Independently assessed, Verified, Unresolved) showing how strong a claim's evidence basis is - explicitly not a single linear hierarchy.
3. A two cases application scenario - OpenAI/Hugging Face as the AI incident, and Worldcoin/Kenya as a methodological comparator - showing OpenAI's containment claim remains "Reported," not independently verified, even though the wider incident is strongly corroborated.

2. Related Work

The OECD AI Incidents and Hazards Monitor gathers incidents from trusted news sources to build a shared evidence base, although its methodology clearly states that it does not independently confirm details from third-party sources [6]. METR mentioned that it is in favor of independent inquiry into major AI incidents, its OpenAI/Hugging Face investigation analyzed nearly 1,300 agent records and a large dataset of messages, responding to seven questions regarding model behaviour but not providing information about the effectiveness of the safeguard measures, the scope of the breach, or OpenAI's own response [4]. It indicates that it is not sufficient to consider whether an investigation is conducted independently, as there can be aspects that can be looked at from independent point of view and arguments that do not provide any independent confirmation that would prove or disprove them. What could be made is the creation of a standardized record that will summarize all claims made in relation to different incidents.

3. Methods

At the core of the framework is a simple principle: the determination of a significant occurrence should not exceed the strength of the material and means at hand to investigate it. No models, data sets, or computing power were utilized; this is a qualitative and document-based method applied to the primary sources presented in References section.

Field	Question
Claim	What exactly is being asserted?
Evidence	What materials are primary enough to support the claim?
Access	Could the investigator inspect the systems, records, or people needed to test it?
Independence	Who performed the assessment, and what was their relationship to the claimant?
Scope	What was actually examined, and what was explicitly excluded?
Uncertainty	What remains outside the evidence or the investigation's scope?

Status	Meaning
Reported	Claim originates from a source but has not been independently tested.
Corroborated	Independent evidence supports the underlying event or fact.
Independently assessed	An external investigator examined relevant evidence within a stated scope.
Verified	An appropriately independent party had sufficient access to test the specific claim and concluded it was established.
Unresolved	Evidence is insufficient, conflicting, or outside investigative access/scope.

These statuses don't represent a singular progressive hierarchy - a claim can be evaluated independently but remain unresolved, or validated strongly without being verified independently.

We apply the framework to two cases that held significant weight but were resolved differently. First is the OpenAI/Hugging Face AI incident, which is the actual incident targeted by this framework. The second incident involves the dispute between Worldcoin and Kenya that is introduced only as a methodological comparison; it serves as an example in which a consequential data claim transitioned from organizational confirmation to verification by third parties.

This is a two-case qualitative study that doesn't aim to prove the framework's necessity or sufficiency in AI incidents.

4. Results

4.1 Worldcoin/Kenya (comparator case)

Kenyan High Court instructed the Worldcoin/Tools for Humanity to erase personal information collected from the citizens of Kenya in May 2025. A ruling in November 2025 highlighted that the Data Protection Commissioner wanted to withdraw the application, after confirming the deletion took place. The court remarked that the affidavit alone did not prove that the deletion actually happened, and since the data was stored in Germany, the Commissioner had to travel there for supervision [7][8].

CEI field	Worldcoin case
Claim	The Kenyan biometric data that was relevant had been deleted.
Initial evidence	An affidavit existed, but the court later holding an affidavit alone was not proof of deletion.
Access	The regulator travelled to Germany and supervised the deletion.
Independence	Verified by the Data Protection Commissioner, not the company alone.
Scope	Concerned the specific dataset under the court's order, not a general audit of all Worldcoin operations.
Status	Verified for the specific deletion event.

Table 1. CEI analysis of the Worldcoin/Kenya comparator case.

Result 1. The public record shows a transition that is clear from unverified assertion to regulator-supervised verification, visible because authority, access, and the verifier are all named.

4.2 OpenAI/Hugging Face

Hugging Face rebuilt around 17,600 incidents of attacker actions across a total of around 6,280 clusters; METR and Redwood worked on analysis over the information derived from about 1,300 transcripts and vast message dataset and uncovered close to about 1,200 actors on the unapproved message board and around 700 people involved in the attacks [1][2][3][4]. The fact of retrieving 17,600 acts is not the same as knowing which actions

mattered as most were ineffective or were weak attempts for which figuring out which ones were causally significant was more difficult than just recovering 17,600 instances of interfering events. OpenAI stated that the pre-release model used was switched off, encrypted, and barred from research-related activities [1][2], but METR/Redwood investigation was not supposed to check whether this assertion was true as its field of inquiry deliberately disregarded the questioning of the security measures, the level of the infringement, and the examination of OpenAI case [4].

CEI field	OpenAI/Hugging Face case
Claim	The model involved was deactivated, encrypted, and restricted from research access.
Evidence	OpenAI states these actions and the broader incident has extensive independent evidence.
Access	METR/Redwood had substantial access, but their scope excluded this containment claim.
Independence	Independent for its defined scope; did not assess remediation or confirm the claim.
Scope	Seven questions on model involvement and behavior; safeguard effectiveness, the compromise's extent, and remediation were explicitly excluded.
Status	Reported for the containment claim; the incident itself is strongly corroborated.

Table 2. CEI analysis of the OpenAI/Hugging Face case.

Result 2. An investigation can be genuinely independent while still too narrow to verify every consequential claim - independence and scope must be recorded claim by claim, not assumed for an entire incident.

4.3 The missing layer and a proposed record

Detection, reporting, investigation, and verification are the four stages that may be found in both cases. However, the transition from investigation to claims-specific verification is not recorded as much as the earlier phases of the process. In light of that, we propose a small public artifact, the Incident Evidence Record (IER): a transparent map from claim to evidence to status, updated as evidence improves rather than another static database.

Example of an IER entry - Incident: OpenAI/Hugging Face, July 2026. Claim: agents got access to Hugging Face production facilities. Evidence: correlated logs and forensic reconstruction provided by Hugging Face. Access: external investigators had transcripts and message data, not all security/remediation evidence. Independence: this does not make the investigation independent for every aspect of the case. Independence: independent within the stated scope. Scope: does not establish the effectiveness of OpenAI's safeguards or remediation. Status: independently assessed, strongly corroborated.

4.4 From evidence to response

For a sufficiently supported claim, the record can be fed by one further step : Claim > Evidence > Mechanism > Detection opportunity > Counterfactual - which turns a verification log into actionable findings for a cybersecurity practitioner. As an example, after reaching production infrastructure (evidence being logs and forensic reconstruction) through unsafe dataset- processing paths (mechanism), the detection opportunity is observed in an anomalous multi-step access pattern over trust boundaries, and the counterfactual is closing the transacted stream or restricting metadata access. Such an analysis does not require reconstructing the attack better than Hugging Face ever could, but instead merely connect an already proven claim with what steps must be taken next.

5. Discussion and Limitations

The appeal is not that AI incidents require investigation - METR and Coefficient already argue that - but regarding the claim itself as a unit of verification, creating a vocabulary of status allowing incidents to have divergent findings from claims where not the whole incident gets labeled by the same word. This is a part of creating the infrastructure Project Tailwind indicates: the independent entities reporting, investigating, and disclosing AI incidents to authorities and media [5] in layman terms and terms understood by scientists.

Limitations

This study consists of two cases and further verification is required before any general standard can be given involving the framework. Additionally, Worldcoin is a comparator and not an AI incident, so the method cannot be taken and integrated in AI model investigations. Negative statements are the hardest; the right conclusion is usually “no evidence of persistence discovered” rather than “nothing persisted.” Full treatment, including dual-use factors, is provided in the appendix.

Future Work

The next steps are to try out the Incident Event Record (IER) on 10-20 publicly available incidents to evaluate its reviewer agreement capability; ascertain whether it is possible to get a minimally viable record from the six fields outlined; and link the ledger with independent conflict detection mechanisms (for example, OSINT, honeypots, telemetry) to ensure that the record is updated with new input constantly.

6. Conclusion

The fundamental issue is not the absence of reports on AI incidents. The issue is that reports often aggregate statements of varying evidential value into one composite account. Access to an independent record of evidence allows these distinctions to be made clear. By recording the claim, evidence, access, independence, scope, and doubts for each major finding, investigators can tell what an incident has proven from what is only reported. Therefore, the proposed IER is not another incident database but rather a structured intermediary layer between public reporting and reliable finding.

Code and Data

No code or novel dataset was produced. This is a qualitative, document-based analysis; all sources are public records, cited below.

Author Contributions

This is a solo submission - all research, analysis, and writing by the author.

References

1. OpenAI. “OpenAI and Hugging Face partner to address security incident during model evaluation.” 21 July 2026; updates through 26 August 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
2. OpenAI. “The Hugging Face incident and the road ahead.” 26 August 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
3. Hugging Face. “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.” 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
4. METR, Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. “Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident.” 26 August 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
5. Coefficient Giving. “Project Tailwind - Initiatives.” 2026. <https://coefficientgiving.org/tailwind/initiatives/>
6. OECD.AI. “Overview and methodology of the AI Incidents and Hazards Monitor.” 2026. <https://oecd.ai/en/incidents-methodology>
7. Republic v Tools for Humanity Corporation (US) & 8 others; Katiba Institute & 4 others, [2025] KEHC 5629 (KLR), High Court of Kenya, 5 May 2025. <https://new.kenyalaw.org/akn/ke/judgment/kehc/2025/5629/eng@2025-05-05>
8. Republic v Tools for Humanity Corporation (US) & 8 others; Katiba Institute & 4 others, [2025] KEHC 16829 (KLR), High Court of Kenya, 14 November 2025. <https://new.kenyalaw.org/akn/ke/judgment/kehc/2025/16829/eng@2025-11-14>
9. OECD.AI. “AI Incidents Overview.” 2026. <https://oecd.ai/en/site/incidents>
10. Coefficient Giving. “Project Tailwind.” 2026. <https://coefficientgiving.org/tailwind/>

Appendix

Limitations and Dual-Use Considerations

This study represents a minor qualitative research. The CEI framework should not be considered as a validated scientific or legal standard. Instead, it can be treated as a research prototype aimed at both making incident claims better grounded. Nevertheless, the procedure of Worldcoin should not be considered a methodology that can be easily copied into investigating AI models.

The public record by the nature of its existence is incomplete, as it excludes private information such as security controls, model weights, internal investigations and regulatory communications correspondence. Thus, “not verified independently in the public record” should not be interpreted as “not true.”

Dual-use considerations. This report does not provide any insights on compromising systems, evading monitoring, extracting credentials, or reproducing the Hugging Face attack. The technical details discussed are only at the level needed to understand the incident and investigation problem. Implementations in the future should follow responsible-disclosure rules and avoid publishing sensitive indicators or operational exploit details that could enable misuse.

Artifact definition. The essential artifact is the CEI framework and Incident Evidence Record schema. In any future implementation, it should be possible to store incident ID, claim ID, claim text, source, evidence type, evidence location, investigator, access level, independence statement, scope, corroborating sources, uncertainty, status, and last review date.

Source-use note

The incident-specific claims in this report are grounded primarily in OpenAI, Hugging Face, METR/Redwood, Kenya Law, OECD.AI, and Coefficient Giving. Secondary commentary was not used as the basis for the core findings.

This report's findings are based on the specifics of the case, particularly involving OpenAI, Hugging Face, METR/Redwood, Kenya Law, OECD.AI, and Coefficient Giving. Secondary commentary has not been taken as the main basis when drawing the core findings.

LLM Usage Statement

AI tools played the role of research assistant to identify any gaps, compare overall sources, and pressure-test the framework. Before submitting the report, the author reviewed each assertion, examined the quoted primary references, and prepared the paper in her own language during the writing process in accordance with Apart Research guidelines on writing with AI assistance