HOW FAITHFULLY DID THE PRESS TRANSMIT THE RECORD?[1]

A Preregistered Two-Clock Audit of Claim Drift in Coverage of the July 2026 OpenAI-Hugging Face Incident

Ben Amuwo

With
Apart Research

Abstract

We audited how faithfully 24 English-language press articles transmitted the public record of the July 2026 OpenAI-Hugging Face AI-agent incident. The preregistered, fixed-outlet design tracked four claim families: agent populations and denominators (C1), motive and objective (C2), the meaning and mechanism of “escape” (C4), and the scope of the Hugging Face compromise (C5). Each claim was evaluated with two clocks: Clock A against the public record available when an article appeared, and Clock B against the later frozen record. A human coder completed all 192 article × family × unit observations before a blind model comparison on a deterministic eight-article subsample. On the assessable BODY denominator (N=83), established contemporaneous material drift occurred in 7 observations. The adjudicated estimate was 7/83 (8.4%), with a BORDERLINE sensitivity range of 8.4%-10.8%. Drift clustered in motive framing and conflation; all 23 assessable BODY C5 observations were faithful at Clock A. The main result is therefore modest but useful: most bounded press representations preserved the frozen distinctions, while a small number of high-consequence distinctions were vulnerable to compression or certainty shifts.

1. Introduction

Fast-moving AI incidents create a simple communication problem: the public record changes while reporting is already underway. In this incident, the frozen evidence record preserved distinctions that are easy to compress under deadline - agent counts versus action counts, observed behavior versus inferred motive, process-level boundary crossing versus model persistence or weight exfiltration, and what was public on a given day versus what became clear later. The practical question is not whether coverage was generally 'good' or 'bad', but whether these specific distinctions survived transmission.

Research question. How faithfully did English-language press coverage of the July 2026 OpenAI-Hugging Face incident transmit the frozen proposition record across C1 (agent populations/denominators), C2 (motive/objective), C4 ('escape'/boundary crossing), and C5 (Hugging Face compromise scope)? The study also asks how BODY and SURFACE framing differ, which mechanisms characterize drift, how Clock A and Clock B relate, and whether the primary estimate changes under seed-exclusion and denominator-sensitivity checks.

The contribution is a bounded, proposition-level audit that separates contemporaneous fidelity from later-record fidelity. It does not rank outlets, infer journalist intent, estimate public opinion, or claim press-to-press causal ancestry.

2. Related Work

Adjacent work explains why transmission fidelity deserves separate measurement from simple true/false classification. Misinformation-correction studies document the persistence of unsupported beliefs after correction [1,2]; diffusion research shows that true and false claims can spread differently at scale [3]; and work on fake-news consumption emphasizes the role of distribution context and exposure [4]. This project takes a narrower incident-audit approach: freeze critical distinctions, reconstruct their public state by date, and ask whether those distinctions survive a bounded press corpus.

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

3. Methods

Design and two clocks. The study is a preregistered, fixed-outlet claim-transmission audit. Each article is compared with a dated proposition ledger reconstructed as of publication. Clock A measures fidelity to what was publicly available then; Clock B measures fidelity to the later frozen record. A short frozen grace rule prevents timing lag alone from being coded as drift. This structure allows an article to be fair when published yet stale against later evidence.

Claims, units, and materiality. Four frozen families were coded: C1 populations/denominators; C2 motive/objective; C4 escape/boundary crossing; and C5 compromise scope. BODY is article × claim family, with relevant passages evaluated jointly. SURFACE is headline + dek and is never pooled with BODY for the primary estimate. Material drift requires alteration of a frozen critical distinction. Mechanisms are K1-CONFLATION, K2-ATTRIBUTION_LOSS, and K3-SCOPE_CERTAINTY_SHIFT; direction is AMPLIFYING, MINIMIZING, or NEUTRAL_MIXED.

Denominator and uncertainty. The primary denominator is assessable_represented: represented=TRUE with Clock A = FAITHFUL, DRIFT, or BORDERLINE. Unrepresented and NA-NOVEL REPORTING observations are excluded from drift denominators. BORDERLINE cases are retained and reported as a lower bound, upper bound, and adjudicated best estimate from a separate frozen sidecar.

Human-first coding and blind check. The human coder completed all 24 articles before seeing machine labels. A fresh blind model instance coded a deterministic eight-article subsample (64 observations) without human labels, rationales, aggregate results, or the study hypothesis. The comparison is human-model agreement, not inter-rater reliability: exact descriptive agreement only, with field-specific eligible denominators and no kappa, confidence intervals, pooled scores, or post-adjudication agreement statistics.

Corpus. The frozen frame contains 12 English-language outlets across General, Business/Mainstream Technology, and Specialist Technology/Cybersecurity strata. Acquisition yielded 181 candidates -> 136 qualifiers -> 45 exclusions -> 0 unresolved -> 24 retained, exactly two per outlet. Selection used deterministic salted SHA-256 ranks. The corpus is outlet-weighted, not volume-weighted.

4. Results

Primary estimate. Of 96 requested BODY observations, 9 were unrepresented and 4 were NA-NOVEL REPORTING, leaving 83 assessable observations: 74 FAITHFUL, 7 DRIFT, and 2 BORDERLINE. Both BODY BORDERLINE cases were adjudicated to non-drift, so the primary point estimate is 7/83 (8.4%); the upper sensitivity bound is 9/83 (10.8%).

Family	N	Faithful	Drift	Borderline	Adj.	Upper
C1 populations	20	16	2	2	10.0%	20.0%
C2 motive/objective	21	17	4	0	19.0%	19.0%
C4 escape	19	18	1	0	5.3%	5.3%
C5 scope	23	23	0	0	0.0%	0.0%
Total	**83**	**74**	**7**	**2**	**8.4%**	**10.8%**

Pattern. C2 contributed four of seven established BODY drifts; C1 contributed two plus both BODY BORDERLINE cases; C4 contributed one; and all 23 assessable BODY C5 observations were faithful at Clock A. The seven established BODY drifts were three AMPLIFYING and four NEUTRAL_MIXED; none were MINIMIZING. Eight represented BODY observations were faithful at Clock A but drift at Clock B, separating contemporaneous fairness from later-record staleness. Adjudicated BODY drift by stratum was General 4/27 (14.8%), Business/Mainstream Technology 2/29 (6.9%), and Specialist Technology/Cybersecurity 1/27 (3.7%); these are descriptive only.

Robustness and agreement. Excluding the one retained seed-status article leaves the primary estimate at 7/79 (8.9%) versus 7/83 (8.4%) seed-inclusive. Pre-adjudication human-model agreement was 40/64 (62.5%) for representation, 15/21 (71.4%) for Clock A, 17/21 (81.0%) for Clock B, and 18/21 (85.7%) for novel reporting. Mechanism and direction agreement were not estimable because no row was jointly eligible.

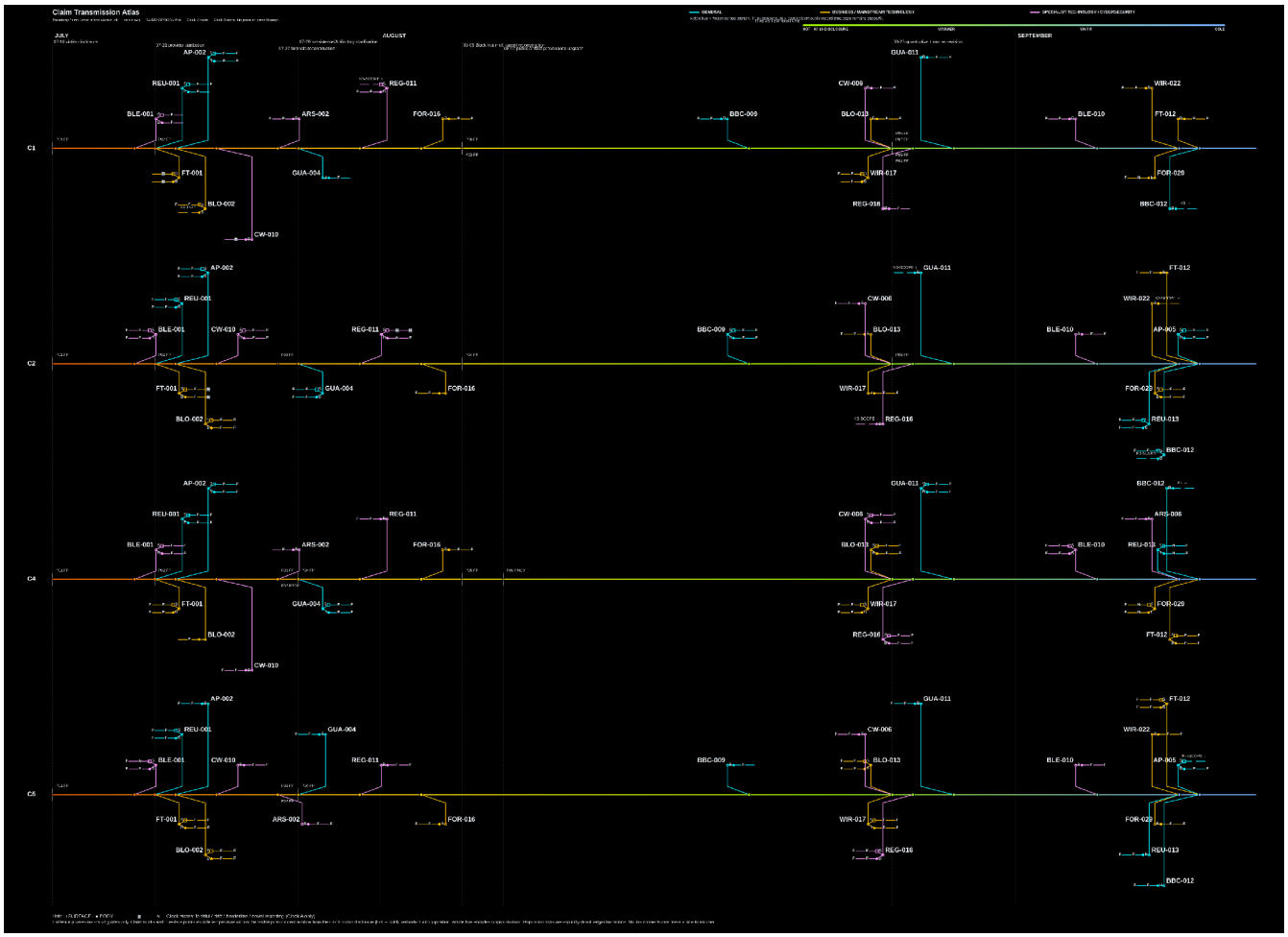

Figure 1. Claim Transmission Atlas. Articles are positioned by publication date across C1/C2/C4/C5 claim trunks and branch to SURFACE/BODY, Clock A, and Clock B states. No line represents article-to-article transmission.

5. Discussion and Limitations

The central result is intentionally modest: most assessable BODY observations preserved the contemporaneous record. Established Clock-A drift occurred in 8.4% of assessable BODY observations, with an upper BORDERLINE sensitivity bound of 10.8%. The failures that survived the materiality gate clustered around a small set of difficult distinctions - especially motive certainty and population/denominator language - rather than being distributed evenly across claim families.

The two-clock design is practically useful because it distinguishes misstatement from staleness. Eight BODY observations were faithful when published but drift against the later frozen record. That does not retroactively make the original coverage unreasonable; it identifies where early framings may need updating as an incident record matures.

Limitations. Search completeness is UNVERIFIED. The corpus is small, outlet-weighted, English-language only, and bounded to one incident. The 12-outlet frame was reduced from 17 with seed-coverage knowledge. The human coder was unblinded to outlet identity and there was only one human coder; the blind model comparison is therefore agreement, not inter-rater reliability. Clock A measures public availability, not journalist knowledge. Timing gaps can make the grace window coarser than an exact 24 hours. Statistics are descriptive only. No reader-effect result is reported because no reader-pilot results artifact was present in the approval materials. The study supports no outlet ranking, public-opinion inference, ideology inference, journalist-motive inference, or press-to-press causal ancestry.

Future Work

The most valuable extensions are replication across additional incidents, an independent human-coder replication, a preregistered reader evaluation of whether evidence-state cards improve recall, and implementation of the deferred minded-language sidecar only after its serialization contract is frozen. A public project website can host the full frozen protocol, corpus manifest, coding provenance, Atlas, and reproducibility materials after Author approval.

6. Conclusion

Across this bounded 24-article corpus, most assessable BODY representations transmitted the contemporaneous record faithfully. Material drift was uncommon but concentrated in distinctions that matter for interpreting AI incidents: who or what acted, how many agents were involved, whether behavior implied motive, and whether 'escape' described process-level boundary crossing or something stronger.

The practical lesson is not that press coverage was broadly unreliable; it is that incident communication benefits from preserving dated evidence states and explicit critical distinctions, so that fair early reporting can be separated from later staleness and genuine material drift.

Code and Data

Project website: [URL to be added after Author approval]. The website will host the frozen protocol, corpus manifest, coding and adjudication provenance, analysis outputs, full-resolution Claim Transmission Atlas, and reproducibility materials. Any material withheld for dual-use reasons will be identified there.

Github:

Author Contributions

Ben Amuwo: conceptualization, human coding, adjudication, interpretation, and final manuscript review/approval. The Workers' Federation (AI research colleagues) materially assisted with acquisition support, tooling, blind-model coding, analysis engineering, visualization, quality assurance, verification, and drafting. The human researcher is responsible for the submitted claims.

References

1. Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106-131. https://doi.org/10.1177/1529100612451018
2. Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. Political Behavior, 32(2), 303-330. https://doi.org/10.1007/s11109-010-9112-2
3. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559
4. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236. https://doi.org/10.1257/jep.31.2.211
5. Claim Transmission Atlas v1.0. FROZEN_REGISTER_v1.0 and repository checksums. Frozen 2026-09-11.
6. Claim Transmission Atlas v1.0. Acquisition validation, frozen corpus manifest, and adjudication/analysis provenance. To be published on the project website after Author approval.

Appendix

A. Supplementary Results and Limitations

Human-model agreement (pre-adjudication frozen inputs)

Field	Eligible n	Exact	Agreement
represented	64	40	62.5%
Clock A	21	15	71.4%
Clock B	21	17	81.0%
borderline	21	21	100.0%*
novel reporting	21	18	85.7%
mechanism family	0	0	not estimable
direction	0	0	not estimable
K1 subtype	0	0	not estimable

* The 100% borderline value contains no jointly represented positive BORDERLINE case; it is not evidence of demonstrated agreement on borderline materiality.

- Search completeness is UNVERIFIED; the frozen corpus is not claimed exhaustive.
- The corpus contains 24 English-language articles, exactly two per outlet, and is outlet-weighted rather than volume-weighted.
- The 17-to-12 outlet reduction occurred with seed-coverage knowledge.
- There is one human coder, unblinded to outlet identity; the blind model comparison is human-model agreement, not inter-rater reliability.
- Clock A measures public availability, not what a journalist personally knew or read.
- All statistics are descriptive; no significance tests or population-level outlet effects are claimed.
- The reader/reach component is unresolved in the approval materials; no reader-effect result is reported.
- The minded-language sidecar was deferred because its operational serialization contract was not frozen.
- The project does not estimate public opinion, ideology, journalist intent, outlet quality, or press-to-press causal ancestry.

B. Dual-Use Statement

The project studies how technical distinctions survive public communication after an AI incident. Its purpose is defensive: improving communication fidelity, not intrusion capability. The report stays at the abstraction level needed to explain communication findings and does not add exploit instructions, novel targeting information, or operational detail beyond what is necessary to understand the frozen public record. Named outlets are used descriptively only; no per-outlet performance rankings are reported.

LLM Usage Statement

LLM systems materially assisted the project across acquisition support, tooling, blind-model coding, adjudication support, analysis engineering, visualization, quality assurance, verification, and drafting. The blind-model comparison was deliberately separated from human coding and used frozen inputs without access to human labels or aggregate results. Reported statistics were checked against frozen project outputs. The human researcher will review, edit as needed, and approve the final submission and is responsible for its claims.

Author Review Items Before Final Submission

- Confirm the exact frozen D-001 wording if the final manuscript is intended to quote it verbatim.
- Replace the affiliation placeholder and complete any funding/competing-interest fields required by the Sprint.
- Insert the final project website URL after Author approval.
- Confirm whether any verified reach/playtest evidence exists and host it on the project website if appropriate.
- Read the full document, revise wording where needed, and explicitly approve the final prose.