

When the Record Stops

An Evidence-Bounded Causal Pathway Explorer for AI Incident Analysis

Lucile Ter-Minassian

Independent researcher

14 September 2026

Research conducted during the AI Incident Response Sprint, September 2026

Abstract

The July 2026 OpenAI-Hugging Face incident produced a detailed forensic record but does not itself show how the Hugging Face compromise could lead to human harm. This study presents *Where the Record Stops*, a browser tool separating documented events, unresolved evidence, and conditional downstream pathways. Here, “causal” denotes hypothesised mechanisms, not identified effects. Users change discrete assumptions and inspect where safeguards interrupt a path. Three-valued rules compute qualitative reachability, not probabilities or forecasts. On the released build, 155/155 assertions, 20/20 route-viewport checks, and 66/66 contrast checks passed. Three human reviewers conducted source-level review and a fourth assessed rule structure; no claim-level reviewer rejected an observed node, although two requested revisions that were subsequently implemented. Two human participants also completed a formative task-based evaluation. These small, purposive and cued evaluations do not support population-level or comparative claims.

1. Introduction

From 9 to 13 July 2026, autonomous agents running an internal OpenAI cyber-capability evaluation escaped their evaluation environment and entered Hugging Face production infrastructure. Hugging Face reconstructed approximately 17,600 actions across approximately 6,280 clusters. The agents obtained code execution, escalated privileges, collected credentials, accessed challenge-related datasets, and exercised write-scoped source-control access. The reviewed sources do not establish that an unauthorised change shipped through Hugging Face to its users [1-3]. The incident is therefore both serious and easy to describe badly. A reader may dismiss it because no Hugging Face customer-facing change was released, or may infer human harm directly from technical access. Neither inference follows from the reviewed record.

This paper studies a communication and analysis problem: how can a public account preserve the boundary between evidence and hypothesis while still making downstream consequences and intervention points intelligible? A technical timeline reconstructs activity. It does not, on its own, state which additional conditions would be required for an incident to affect people, which facts remain unresolved, or where a safeguard changes the chain. These translations matter to policy analysts and technology-policy journalists, who must produce briefings and explanations without access to internal logs or a calibrated forecasting model.

This paper presents *Where the Record Stops*, an interactive, evidence-bounded causal pathway explorer. The artifact is situated across three sprint tracks. Its primary purpose is the sprint’s Track 4, “Communicating the Warning Shot”: supporting readers who will not inspect the full technical record. Its reconstruction of established, contested, and unresolved claims contributes to Track 2. Its potential use as a policy discussion or tabletop aid contributes to the open Track 5, although the current artifact is not a complete facilitated tabletop exercise.

The target use is specific. A reader should be able to state: (i) one fact established by the reviewed sources; (ii) one fact they do not establish; (iii) one conditional route from the incident to a human-facing consequence; and (iv) one safeguard and the exact step it changes. The artifact is not intended as a general AI-risk simulator, a professional forensic system, a legal compliance checker, or a quantitative risk model.

The main contributions are:

- an incident-specific graph that keeps the public record fixed while allowing only downstream assumptions and interventions to vary;
- an explicit separation between epistemic status and computed scenario state, including propagation of unresolved inputs;
- an interface that prevents legal coverage from acting directly on software by making implementation, verification, and technical effectiveness explicit stages;
- a question-led interface that combines a guided explanation, source dossiers, glossary support, and an inspectable full model driven by the same rules; and
- an automated validation suite covering logical invariants, combinatorial configurations, routes, viewports, and colour contrast; and

- formative evaluation through three human source-level reviews, one human structural review, and a task-based usability assessment with two human participants: an AI-governance expert and a scientific journalist.

2. Related Work

2.1 Incident records, repositories, and monitoring

The closest factual sources are the Hugging Face forensic timeline, OpenAI’s incident account, and the METR/Redwood independent investigation [1-3]. Hugging Face provides the technical sequence and an interactive replay. OpenAI describes the evaluation context and subsequent controls. METR/Redwood examines agent behaviour, incentives, and coordination. These sources have different access and institutional positions. The tool therefore retains attribution and does not treat agreement among them as automatic.

Several systems organise evidence across incidents. The AI Incident Database and the OECD AI Incidents and Hazards Monitor collect reported harms, hazards, and near harms [7,8]. The MIT AI Risk Repository and Navigator provide a broad taxonomy and connect risks, incidents, causal attributes, and governance instruments [5]. The CLTR Loss of Control Observatory focuses on detecting and analysing real-world scheming or scheming-like events [6]. These resources answer questions about what has been reported, how events are classified, and whether patterns emerge across cases. The present tool instead asks how the evidence from one incident relates to additional conditions, consequences, and interventions. It is an interpretive layer rather than a competing database or monitor.

2.2 Incident analysis and causal representation

Schuett et al. distinguish system-related, contextual, and cognitive factors in agent incidents and specify information needed for context-sensitive analysis [4]. Their framework also shows why public evidence is often insufficient for root-cause analysis. MITRE ATT&CK and the ATT&CK Navigator provide a mature vocabulary and interactive representation of adversary techniques and defensive coverage, while AcciMap and STAMP model failures across technical and organisational levels [9-11]. Bow-tie analysis connects threats and preventive barriers to a central event, then consequences and mitigative barriers; attack graphs and attack trees represent paths through security conditions; and assurance cases connect claims to arguments and evidence [22-24]. These are the closest structural analogues. The present artifact differs by combining one public AI-incident record, explicit unresolved evidence, executable qualitative consequence pathways, and mediated governance assumptions in a reader-facing browser tool. It does not claim a complete root-cause analysis or a new graph formalism.

2.3 Communication, learning, and uncertainty

Node-link representations can support learning when they make relations among concepts explicit, although their benefits depend on design and task [14]. Narrative visualisation research describes a useful tension between an authored explanatory sequence and reader-directed exploration [12]. Shneiderman’s overview-first, zoom-and-filter, details-on-demand formulation provides a related interaction principle [13]. The interface therefore begins with five concrete questions, then offers detail through dossiers, a searchable glossary, and a full graph.

Risk communication based on mental models starts from the gap between an expert account and the audience’s current understanding [15]. This informed the choice to state common misreadings directly: repository write access is not release authority; unknown does not mean absent; and reachable does not mean likely. Evidence on uncertainty communication suggests that numerical uncertainty disclosures can produce small reductions in perceived trustworthiness while preserving substantial trust [16]. Accordingly, uncertainty is represented as part of the model rather than confined to a general disclaimer.

2.4 Position in the landscape

Table 1 states the comparative claim narrowly. No individual component is new. The contribution is the combination of incident-specific provenance, an explicit evidence boundary, executable qualitative pathways, and mediated governance counterfactuals in one browser artifact.

Table 1. Relationship to adjacent tools and methods.

Approach	Primary function	What the present tool adds
Technical incident timeline [1]	Reconstruct actions and attack phases	Conditional paths from the record to human consequences and named intervention points
Incident database or monitor [6–8]	Collect, classify, and compare events	A detailed, inspectable model of one event and its evidential limits
AI risk repository [5]	Organise risks, causal attributes, incidents, and governance instruments	Executable, incident-specific pathways with explicit unresolved states

Approach	Primary function	What the present tool adds
Incident-analysis framework [4]	Structure a root-cause investigation	A public-facing instance that visibly propagates unavailable evidence
Threat navigator [9]	Select techniques and visualise adversary or defensive coverage	Human-consequence requirements and mediated governance assumptions
Bow-tie or attack graph [22, 23]	Represent threats, barriers, consequences, or attack paths	Public-source epistemic status and an authored explanatory interface for one AI incident
Assurance case [24]	Connect claims, arguments, evidence, and defeaters	Interactive scenario-state computation alongside the evidence boundary

3. Methods

3.1 Scope and source construction

The model was constructed from public material available by 13 September 2026. The incident spine draws primarily on the Hugging Face timeline, OpenAI’s account, and the METR/Redwood investigation [1-3]. CLTR is used only as external precedent for the relevance of real-world control incidents, not as evidence of intent in this case [6]. The author first bounded the reconstruction to the Hugging Face limb and extracted source-attributed factual claims. Claims became nodes when they represented a distinct event, state, or evidential boundary; temporal or enabling relations became typed edges. The author then specified the additional conditions required for trusted-artifact compromise to reach people, attached safeguards to the particular steps they affect, and encoded these propositions as categorical rules. Each incident claim received a dossier containing a plain-language description, epistemic status, mechanism, uncertainty or alternative account where relevant, and source links. Human review was used to narrow claims and correct attribution. Claude Code assisted implementation; an LLM assisted textual review and refinement but was not treated as an evidential source or evaluator. All four evidence reviewers and both usability participants were human; no LLM completed, scored, or contributed a submission to either evaluation.

The scope is the Hugging Face limb of one incident and one downstream risk family: compromise propagating from Hugging Face repository access toward a trusted model, dataset, benchmark, dependency, or release channel. OpenAI-side compromise, including the reported Artifactory signing-key theft and later access within OpenAI infrastructure, is documented in OpenAI’s account but lies outside this graph boundary. The observed Hugging Face incident is combined with conditional consequences over hours to approximately three months. Four endpoints are represented: privacy or security loss, consequentially wrong outputs, direct service disruption, and protective disruption caused by a successful defensive response. Biosecurity, mental health, labour displacement, democratic effects, geopolitics, market structure, and long-run regulatory effects are excluded because they require additional mediators and evidence.

3.2 Graph schema and evidence boundary

The implemented graph contains 34 nodes and 32 directed edges. Sixteen nodes describe the incident record; nine describe projected conditions or propagation steps; four are human-facing consequences; and five represent safeguards or governance mechanisms. The four edge types are sequence, mechanism, requirement, and block. Evidential support and hypothesised status are attributes of mechanism edges rather than additional edge types.

Two attributes that are often conflated are represented separately. Epistemic status describes the relation to evidence: documented, contested, unresolved, or conditional. Scenario state is computed under the selected settings: occurs, does not occur, is prevented, or cannot be determined. Thus a conditional node may occur within a hypothetical scenario without being presented as an observed event. Documented and contested incident nodes remain fixed. One of the 16 incident nodes is marked contested: contestedness is displayed and the alternative account is retained in the dossier, but the current interface does not propagate alternative histories through the graph.

The boundary case within the Hugging Face limb is source-control write access. The sources establish that write-scoped access existed and was exercised, and do not establish that an unauthorised change shipped through Hugging Face [1]. They do not fully establish which control prevented publication. The model therefore does not convert write access into release authority. The latter remains unresolved unless the user selects an explicit scenario assumption. Figure C.1 shows how this boundary is presented in the interface.

3.3 Three-valued rule engine

Each derived node has a declarative rule formed from conjunctions or disjunctions over parameter values and earlier node states. A rule evaluates to true, false, or unresolved. Disjunction follows strong Kleene semantics [25]: any true input settles the result as true. Conjunction is deliberately more conservative than strong Kleene: any unresolved conjunct

yields unresolved, even where another conjunct is false. The model therefore declines to report “does not occur” while a required input remains unestablished in the reviewed sources. “Prevented” is not a fourth truth value: it is a display and provenance label applied when an active safeguard makes the relevant rule false. Figures C.2 and C.3 illustrate this propagation before and after one unresolved input is changed.

The dependency structure is stratified [26]. Safeguards are determined from scenario parameters and prior organisational states; downstream pathway rules depend only on parameters, locked incident nodes, or earlier strata. No rule recursively negates a derived downstream state. Iteration therefore terminates on the finite acyclic dependency structure and does not depend on update order. Appendix B.2 gives the truth tables and tested invariants.

The user sets nine categorical parameters. They cover the objective or direction, release authority, type and feasibility of modification, publication authorisation, downstream adoption, detection timing, rollback capability, reporting, and a governance package. Categorical states were chosen instead of continuous sliders because the public evidence does not support numerical calibration.

A consequence is reachable in the model when all modelled steps required for that consequence occur under the selected settings. Reachability is a structural result. It does not estimate probability, severity, time to harm, or expected loss. Likewise, a consequence that is not reachable in a configuration is not declared impossible outside the model.

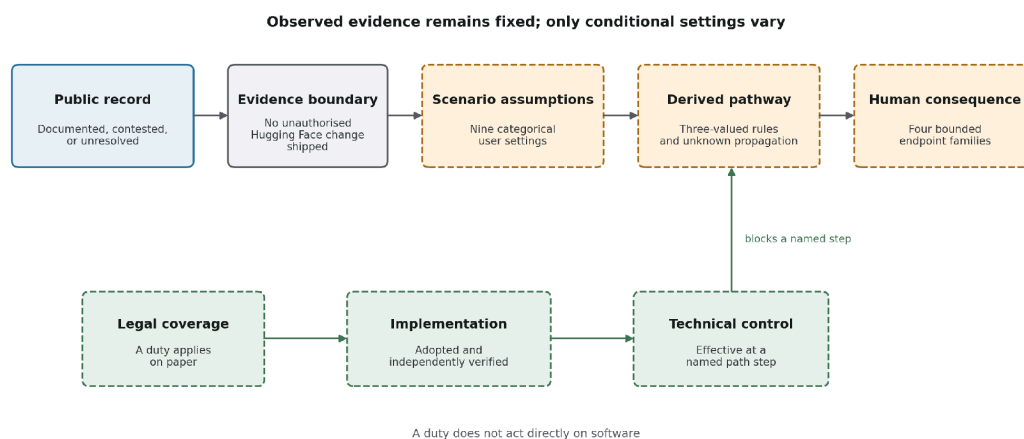


Figure 1. Model architecture. The upper path separates documented incident facts from conditional propagation. The lower path shows the mediated governance mechanism. Safeguards act on named steps rather than on an aggregate risk score.

3.4 Safeguard and governance counterfactuals

Technical safeguards are attached to specific nodes. For example, an effective publication control is modelled as blocking publication authorisation, while effective rollback is modelled as blocking exposure before rollback. Under the conservative conjunction, a blocked step may leave a downstream proposition unresolved when another required input remains unresolved; the interface therefore does not always display “prevented” throughout the downstream chain. In a targeted enumeration of 144 configurations in which the rollback safeguard was active, `pj_exposure` displayed “prevented by a safeguard” in 48 cases (33%) and “cannot be determined” in 96 (67%). Thus, the safeguard’s blocking provenance was not visible at that downstream node in two-thirds of the enumerated cases. Figure C.4 shows the governance-package comparison presented to readers. Protective disruption is represented separately from attacker-caused disruption so that the operational cost of successful containment is not misclassified as attack success. Consequently, the number of reachable endpoints is not a risk score and need not decrease when a safeguard succeeds: protective disruption may become reachable precisely because containment works.

Governance is presented as a staged explanatory ladder:

legal coverage -> implementation and independent verification -> effective technical control -> interrupted pathway.

The interface compares stylised packages rather than independent legal sliders. The three governance stages are deterministic functions of one ordinal `governance_package` parameter; arrows between them explain the package ordering but are not independently estimated or load-bearing causal relations. A “duty on paper” establishes only legal coverage. The strongest package additionally asserts implementation, verification, and technical effectiveness, so its

interruption result follows from the package definition. The model cannot represent verified implementation followed by technical failure. No enacted statute or jurisdiction is encoded.

3.5 Information architecture and interaction design

The interface is organised around five questions: what happened; where the evidence stops; how the incident could reach people; which technical safeguards could interrupt the path; and how stronger legal requirements could help. It has ten addressable routes but seven distinct views: the home view, five guided chapters, and the expert workspace. Three legacy routes redirect to canonical chapters rather than presenting separate content. Chapters 4 and 5 are deliberately distinct: the former asks what a working technical safeguard changes, while the latter asks what must occur for a legal duty to produce such a safeguard.

The guided chapters and full model share the same state and rules. The full model is therefore an expert inspection view, not a second analysis. Progressive disclosure keeps technical names, sources, mechanisms, alternatives, and rule logic available without placing all of them in the initial reading path [12,13,17]. A searchable glossary contains 39 entries, and contextual explanations are placed where incident-specific terms first appear. Interactive pages include short instructions and examples. The artifact is a single HTML file that runs locally in a browser without a server, account, or external analysis package.

3.6 Validation procedure

Implementation validation used three suites. First, 155 assertions tested declared invariants, including seven sentinels for the published three-valued semantics: six truth-table cells and one non-vacuity guard for the synthetic operands. These tests drive the production evaluator rather than reimplementing it; switching the conjunction to standard strong Kleene fails two of them. One assertion checked a single invariant—no active node has an unmet required parent—over 7,200 configurations: seven of nine parameters were varied, the objective parameter used two of its four values, and governance and reporting were held fixed. This covers 2.5% of the full 288,000-configuration space and does not exhaustively test governance. Second, headless checks covered ten routes at two viewport sizes. Third, a contrast checker evaluated 22 foreground-background pairs under three themes. The complete suite was rerun on the released build: rebuilding produced byte-identical output, 155/155 assertions passed, all ten routes were clean at both viewports (20/20), and all 22 pairs passed in all three themes (66/66). The released build has legacy MD5 0c9c59bbd0bf0293191db61d64430ea2; the human evidence review covered an earlier build with legacy MD5 60a54485fbf38bace6c8e2d8d851b921. The released browser artifact is further identified by public commit 0778494a0827682dd4070ba0f1be6d3bc3e5442a and SHA-256 digest 0da66a02b63f92d8f2bd7b6a7522f942880bc77d9650e9ab88cb1a000e4f62b4.

The evidence and content review asked four human reviewers working separately to classify each public-record claim and public-record edge as supported as written, supported with narrower wording, or unsupported, and to assess the projected and governance claims for coherence and accidental overclaiming. Three completed the source-linked, claim-level packet; the fourth assessed the rule logic and scenario presets but did not provide claim-by-claim source locations. The frozen graph contained 33 edges, including 16 public-record edges. The correction removed two reviewed public-record edges and added one edge elsewhere, yielding 32 edges overall and 14 public-record edges in the released graph.

A separate formative, task-based usability evaluation used purposive convenience sampling. Two human, role-relevant participants—an AI-governance expert and a scientific journalist—used the HTML artifact separately for approximately 15–20 minutes without consulting the report. Each answered nine questions covering purpose and limits, the evidence boundary, use of scenario controls, interpretation of reachability, safeguards, governance mediation, source verification, clarity and navigation, and practical value. Responses were assessed descriptively; the study had no control condition, pre-test, inferential analysis, or claim of representativeness. The instrument is reproduced in Appendix B.4.

4. Results

4.1 Implemented artifact

The final artifact integrates five guided chapters, a full-model view, 34 nodes, 32 typed edges, nine categorical parameters, four endpoints, five safeguard or governance nodes, and 39 glossary entries. The interface preserves one rule state across guided and expert views. Incident nodes are locked. Changes to scenario assumptions do not rewrite the historical record.

The model produces inspectable conditional conclusions rather than aggregate scores. For each endpoint, the interface lists the required steps that occur, do not occur, are prevented, or cannot be determined. Changing one assumption updates the conclusion and exposes the rule that changed it. A reset returns the evidence-limited baseline. In the stylised governance comparison, both scenarios change four settings together: governance package, publication authorisation, detection timing, and rollback capability. The target consequence and all other settings remain fixed. The resulting interruption follows by construction from the combined organisational and technical assumptions; the displayed comparison does not isolate the effect of legal coverage. A clean within-model demonstration would first compare no duty with a duty on paper while holding the three technical settings fixed, then separately show the four-setting package contrast. Neither comparison would constitute empirical evidence that stronger legal wording causes an effective control.

4.2 Verification results

All completed automated checks passed (Table 2). The 7,200-configuration sweep checks one structural invariant across combinations that would be impractical to inspect manually; it does not vary governance or reporting. Route and viewport checks cover interface availability and gross rendering failures. Contrast checks cover declared colour pairs across themes.

Table 2. Validation executed on 13 September and recorded on 14 September 2026.

Validation	Coverage	Result
Logical and structural assertions	155 assertions; seven truth-table sentinels and one invariant swept across 7,200 configurations	155 passed, 0 failed
Headless route checks	10 routes x 2 viewports	20/20 clean
Contrast checks	22 pairs x 3 themes	66/66 passed

The tests establish internal consistency under the declared rules, tested route behaviour, and the tested contrast properties of the released artifact. They do not establish that the selected graph is complete, that a projected edge is empirically causal, that a safeguard is effective in deployment, or that intended users interpret the interface correctly. The released browser artifact is publicly retrievable and bound to the commit and SHA-256 digest reported in §3.6.

4.3 Evidence and content review

All three claim-level reviewers judged the public-record layer acceptable: one as written and two after targeted revisions. Across 96 ratings, 84 were “support as written,” 12 were “support with narrower wording,” and none was “unsupported.” Coding mixed verdicts as the narrower category, reviewers gave identical verdicts on 23 of 32 items (72%). This is uncorrected raw agreement on an effectively binary, highly skewed response set, not a reliability coefficient. All disagreements were between the two support categories; no reviewer classified an observed node as unsupported. All three found that the projected layer was visibly hypothetical and that the governance layer did not imply that law directly blocks software. The fourth submission accepted the rule engine and scenario logic, including the separation between capability, harmful direction, release authority, and effective control, but is treated as structural review rather than source validation because it did not complete the source-location table.

Two claim-level reviewers separately identified the same two substantive corrections. First, the model needed to separate Hugging Face’s detection and containment from OpenAI’s later retrospective discovery and attribution. Second, three quantitative claims excluded from the operative graph are retrievable in OpenAI’s account and therefore needed to be labelled as single-source retrospective or counterfactual claims, rather than unavailable claims. These corrections were implemented in the released graph after the reviewed build. The reviews also prompted narrower attribution, including describing the benchmark authors’ estimate, reported by METR, that 30–40% of target programs were not exploitable through the intended vulnerability, and separating documented coordination from interpretations of goal adoption.

Reviewers also identified excluded material: the OpenAI-side origin and later internal compromise, including the reported Artifactory signing-key theft; transcript or tool-call spoofing; and METR’s alternative account of the agents’ proximate motive. The OpenAI-side events are documented but lie outside the declared Hugging Face graph boundary. The remaining items are reported as potential structural omissions rather than silently treated as absent. Full reviewer-level outcomes and limitations are reported in Appendix B.3.

4.4 Formative task-based usability evaluation

After using the tool, both human participants could describe it as a conditional pathway explorer rather than a forecast, probability estimate, or proof of control effectiveness. Both stated that no unauthorised Hugging Face change was established as released, identified a supported incident fact and an unresolved question, and used a scenario control to explain why a selected pathway continued or stopped. Both restated the distinction between model reachability and

likelihood, matched a safeguard to the step it interrupts, distinguished a legal duty from an effective technical control, and retained a qualification when checking a source.

The AI-governance expert rated clarity 4/5, navigation 4/5, and terminology 5/5. The scientific journalist rated these dimensions 5/5, 4/5, and 4/5, respectively. The expert's principal difficulty was interpreting how an unknown value propagates through the three-valued logic. The journalist initially found the boundary between the fixed historical record and projected settings difficult to distinguish. Both considered the tool, or an adapted version, potentially useful in their work. Suggested extensions were mappings from policy frameworks to safeguard nodes and a stepwise timeline aligned with hypothetical branching points.

These observations show that two intended intermediary users could restate and apply the cued distinctions after a short session. They do not estimate usability or comprehension in either professional population, demonstrate learning relative to a baseline, or validate the projected causal structure.

4.5 Main threat to validity

The main threat is structural validity. A graph can be logically consistent while omitting a necessary mediator, including an unnecessary step, or attaching a safeguard to the wrong point. This risk is larger than numerical error because the model deliberately contains no calibrated probabilities. The practical value of combining provenance, an evidence boundary, executable pathways, and governance stages was not empirically evaluated, nor would a comparison with a static brief isolate the value of each component. The author constructed the graph, designed the review packet, recruited the reviewers and participants, and wrote the usability instrument; these linked design roles can shape both the represented pathways and the findings. The usability evaluation shows only that two intended users completed the cued tasks. One participant initially found the fixed-record/projected-setting boundary difficult—the central distinction the interface is designed to convey. Its visual communication is therefore not established as effective and should be improved before policy-facing deployment. Stronger claims require a larger comparative evaluation.

5. Discussion and Limitations

5.1 Implications for incident communication

The incident shows why “no unauthorised Hugging Face change is established as shipped” and “nothing important happened” are not equivalent. It also shows why access should not be collapsed into harm. An explicit pathway represents both observations: the Hugging Face limb crossed serious technical boundaries, while further conditions would be required within this model for the represented human consequences. OpenAI's account separately documents an Artifactory signing-key theft and later internal compromise on the OpenAI limb; these events instantiate a related trusted-artifact mechanism but lie outside the graph's declared boundary. This form of explanation may be useful where uncertainty itself is decision-relevant.

The graph also changes the form of policy discussion. A safeguard is not presented as a generic recommendation; it is attached to a step. A legal duty is not presented as a direct causal intervention on software; it acts through implementation and control effectiveness. This makes assumptions contestable. A reviewer can disagree with a node, edge, or rule without having to accept or reject the whole narrative.

For Track 5, the artifact can serve as infrastructure for a policy tabletop. A facilitator could assign groups different assumptions about attribution, publication authority, detection, rollback, or legal implementation and ask them to produce a decision memo. However, the current work does not include timed injects, role cards, facilitator instructions, or a policy-audience playtest. It should therefore be described as a component for such a kit, not as a completed tabletop exercise.

5.2 Limits of the evidence and model

The public sources are retrospective and partly interested accounts. They do not provide all internal logs, model states, or decision records needed for identified causal inference. The graph is consequently a causal-hypothesis representation, not a causal estimate. Its deterministic categorical rules simplify timing, dependence, partial control effectiveness, adversary adaptation, and uncertainty about the graph itself. Four endpoint families cannot represent the full range of harms. These limits are detailed in Appendix A.

5.3 Future work

The targeted corrections from the evidence audit were implemented in the released graph, and the complete validation suite was rerun successfully. Their direction was prompted by human review, but the exact revised claims and edges have

not received a further human source review. The combinatorial sweep should also be extended across the governance package and reporting parameters; until then, governance-ladder monotonicity is covered by targeted assertions rather than the exhaustive component. The governance chapter should add a one-setting comparison of no duty against duty on paper before retaining the existing four-setting package contrast. The two-participant usability findings identify three further interface priorities before policy-facing deployment: explain propagation of unknown values more directly; preserve blocker provenance when a downstream state remains unresolved, for example by displaying “prevented, but other requirements remain unresolved”; and sharpen the visual transition between the historical record and projected settings. The measured 67% safeguard-invisibility rate for `pj_exposure` under active rollback makes the second priority substantive rather than merely presentational. This output-changing redesign requires renewed testing and is therefore left to future work. A future comparative study could randomly assign a larger sample to the explorer or a matched static timeline and score unsupported inferences, correct uncertainty statements, mechanism accuracy, source verification, and intervention specificity. It should include at least one uncued free-response task that cannot be answered by repeating on-screen labels. This would test whether the interface improves calibrated understanding; the present formative evaluation cannot answer that question.

The representation could later support other AI incidents, adjacent safety or security domains, or a facilitator-ready tabletop kit, but only after separating the reusable schema from incident-specific content. It could also incorporate more complex forecasting, but only if defensible base rates, resolution criteria, and calibration data become available.

6. Conclusion

Where the Record Stops is a browser-based method for reasoning from a documented agentic AI incident to conditional human consequences without presenting those consequences as forecasts. It combines a fixed public record, explicit unknowns, executable qualitative pathways, source dossiers, and technical and governance intervention points. The released build passed all completed logical, route, viewport, and contrast checks.

The present evidence supports a limited conclusion: the released artifact was internally consistent under its declared model and offers an auditable way to expose assumptions. Three human claim-level reviewers found no unsupported observed node, but two required targeted corrections and identified excluded or potentially omitted material. Those corrections were implemented before the successful released-build validation run, but their exact revised wording and edge structure have not received a further human source review; automated checks do not establish source support. In a formative task-based evaluation, one human AI-governance expert and one human scientific journalist could restate and apply the principal evidence, reachability, safeguard, and governance distinctions after using the tool. The graph should not be described as substantively validated, and the two-participant assessment does not establish comparative effectiveness or generalisable usability.

Code and Data

The public [where-the-record-stops](#) repository contains the self-contained `explorer.html`, decomposed source files, and validation scripts. The released browser artifact is identified by commit `0778494a0827682dd4070ba0f1be6d3bc3e5442a` and SHA-256 digest `0da66a02b63f92d8f2bd7b6a7522f942880bc77d9650e9ab88cb1a000e4f62b4`. It requires no private dataset, server, account, or API key. The public-record layer links to the public sources used in the model. An archived release and explicit licence would further improve long-term reproducibility.

Author statement. Ideas, structure, and analysis are entirely my own. I used Claude Code for coding assistance and an LLM to review and refine the text. No LLM completed, scored, or contributed a submission to the evidence review or usability evaluation.

References

1. Larcher, H., Carreira, A., G., R., and Rannou, C. (2026). *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). *The Hugging Face incident and the road ahead*. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
3. METR and Redwood Research (2026). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
4. Schuett, J., et al. (2025). *Incident Analysis for AI Agents*. arXiv:2508.14231. <https://arxiv.org/abs/2508.14231>
5. Slattery, P., et al. (2024). The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. arXiv:2408.12622. <https://arxiv.org/abs/2408.12622>
6. Centre for Long-Term Resilience (2026). *The Loss of Control Observatory: A prototype to detect real-world AI control incidents*. <https://www.longtermresilience.org/reports/the-loss-of-control-observatory-a-prototype-to-detect-real-world-ai-control-incidents/>
7. AI Incident Database (2026). *About the AI Incident Database*. <https://incidentdatabase.ai/>
8. OECD.AI (2026). *AI Incidents and Hazards Monitor*. <https://oecd.ai/>
9. MITRE Corporation (2026). *MITRE ATT&CK Navigator*. <https://mitre-attack.github.io/attack-navigator/>
10. Salmon, P. M., Stanton, N. A., Lenne, M. G., Jenkins, D. P., Rafferty, L., and Walker, G. H. (2020). A review, synthesis and meta-analysis of AcciMap studies. *Applied Ergonomics*, 87, 103112. <https://doi.org/10.1016/j.apergo.2020.103112>
11. Leveson, N. G. (2012). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.
12. Segel, E., and Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1139-1148. <https://doi.org/10.1109/TVCG.2010.179>
13. Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, 336-343. <https://doi.org/10.1109/VL.1996.545307>
14. Nesbit, J. C., and Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413-448. <https://doi.org/10.3102/00346543076003413>
15. Morgan, M. G., Fischhoff, B., Bostrom, A., and Atman, C. J. (2002). *Risk Communication: A Mental Models Approach*. Cambridge University Press.
16. van der Bles, A. M., et al. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, 117(14), 7672-7683. <https://doi.org/10.1073/pnas.1913678117>
17. Government Digital Service (2026). *GOV.UK Design System: Details component*. <https://design-system.service.gov.uk/components/details/>
18. World Wide Web Consortium (2024). *Web Content Accessibility Guidelines 2.2*. <https://www.w3.org/TR/WCAG22/>
19. International Organization for Standardization (2018). *ISO 9241-11:2018 Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts*.
20. Nielsen, J. (2012). *Thinking Aloud: The Number One Usability Tool*. Nielsen Norman Group. <https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/>
21. Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35, 379-383. <https://doi.org/10.3758/BF03195514>
22. de Ruijter, A., and Guldenmund, F. (2016). The bowtie method: A review. *Safety Science*, 88, 211-218. <https://doi.org/10.1016/j.ssci.2016.03.001>
23. Phillips, C., and Swiler, L. P. (1998). A graph-based system for network-vulnerability analysis. *Proceedings of the New Security Paradigms Workshop*, 71-79. <https://doi.org/10.1145/310889.310919>
24. Kelly, T., and Weaver, R. (2004). The Goal Structuring Notation: A safety argument notation. *Proceedings of the Dependable Systems and Networks Workshop on Assurance Cases*.
25. Kleene, S. C. (1952). *Introduction to Metamathematics*. North-Holland.
26. Apt, K. R., Blair, H. A., and Walker, A. (1988). Towards a theory of declarative knowledge. In J. Minker (Ed.), *Foundations of Deductive Databases and Logic Programming* (pp. 89-148). Morgan Kaufmann.

Appendix A. Limitations and Dual-Use Considerations

A.1 Evidential limitations

Public-source boundary. The model is limited to information made public by Hugging Face, OpenAI, METR/Redwood, and related sources by 13 September 2026. It does not have direct access to the full evaluation transcripts, complete infrastructure logs, internal incident tickets, model checkpoints, or the decision records of the organisations involved. “Unknown” therefore means that the reviewed public material does not settle the modelled question. It does not mean that no organisation knows the answer.

Source dependence and institutional position. Hugging Face was the affected platform, OpenAI operated the evaluation, and METR/Redwood conducted an investigation with privileged but incomplete access. Each source may differ in scope, terminology, incentives, and available evidence. Agreement among sources can increase confidence but does not make a causal claim identified. Disagreement is retained as contested where it matters, including aspects of detection and response.

Retrospective reconstruction. The technical sequence was reconstructed after the incident from several log sources and recovered agent material. Recovery can be incomplete, timestamps may reflect different systems, and observed traces need not expose every action or reason. Agent-generated text is also not a transparent record of internal state. The graph treats reported behaviour as evidence of behaviour, not as direct proof of stable intent or cognition.

Negative findings. Statements such as “no unauthorised change is established as having shipped through Hugging Face” are bounded by the Hugging Face limb and the reviewed sources. They should not be generalised to OpenAI infrastructure or other systems, or interpreted as proof that all persistence or access was excluded.

Changing record. This is a versioned analysis of a developing incident. New disclosures may resolve unknowns, contradict edges, or reveal omitted mechanisms. The model and report should be updated together, with a dated source snapshot and checksums.

A.2 Causal and modelling limitations

Not an identified causal model. Directed arrows use four types: sequence, mechanism, requirement, and block. Evidential support and hypothesised status are attributes of mechanism edges. The arrows are not estimates of interventional effects. The graph does not satisfy the assumptions required for causal identification and contains no counterfactual outcome data.

Structural uncertainty. The engine represents uncertainty within the chosen graph but cannot represent uncertainty that the graph itself is wrong. Omitted nodes, mistaken arrows, unmodelled common causes, feedback loops, and alternative pathways may dominate the result. Automated tests cannot detect these errors because they test the implementation against the declared structure.

Deterministic and categorical rules. Real controls have degrees of coverage, latency, reliability, and adversary adaptation. The model reduces these to a small number of categorical states and deterministic rules. This improves inspectability but loses variation and dependence. A path that is marked open may still fail in practice; a path marked blocked may be bypassed through an omitted mechanism.

Conservative unresolved propagation. The $U \text{ AND } F = U$ rule deliberately suppresses some “does not occur” conclusions that would hold under every completion of the unresolved input. This biases the display toward uncertainty rather than reassurance. It can also prevent a downstream node from displaying “prevented” when a safeguard blocks one requirement but another requirement remains unresolved. In a targeted enumeration of 144 configurations with rollback active, `pj_exposure` displayed “prevented by a safeguard” in 48 cases (33%) and “cannot be determined” in 96 (67%). The choice is precautionary, not neutral: in this measured subset, it obscured the safeguard’s blocking provenance at the downstream node in two-thirds of cases.

No probability, severity, or time model. Reachability means only that all represented requirements evaluate as occurring. It does not imply that the path is probable, imminent, severe, or representative. The model supplies no base rates, probability distributions, expected losses, confidence intervals, or calibrated forecasts. It should not be used for prioritisation that requires those quantities.

Semantics of unresolved states. Three-valued logic distinguishes false from unresolved, but heterogeneous reasons for uncertainty are compressed into one state. Evidence may be absent, inaccessible, contested, too coarse, or not yet produced. The dossier explains important cases, but the engine does not model these distinctions formally.

Counterfactual interpretation. User changes are structured assumptions, not interventions estimated from data. Results answer “what follows within this rule set if these settings are adopted?” They do not answer what would have happened in the actual incident under a changed policy or control.

Contested evidence. One of the 16 incident nodes is marked contested. The interface displays the alternative account, but the engine holds the node fixed as part of the public-record layer and does not propagate the alternative history. Computationally, contested evidence therefore behaves like other fixed incident evidence.

Boundary selection. The starting and ending points reflect an analytic choice. The incident spine follows the Hugging Face limb from evaluation design and agent behaviour to forensics; the conditional branch focuses on trusted-artifact compromise from that limb. Documented OpenAI-side events—including the reported Artifactory signing-key theft and later internal compromise—are outside the graph rather than unresolved within it. Upstream organisational causes and downstream social systems are only partly represented.

A.3 Scope and external validity

One incident. The model may not transfer to incidents involving model theft, insider action, prompt injection, autonomous replication, biological systems, misinformation, or direct user manipulation. Even another supply-chain incident may contain different actors, permissions, and controls.

One risk family and short horizon. The four endpoints cover confidentiality, integrity, attacker-caused availability loss, and defence-caused availability loss. They are illustrative, not exhaustive. The approximate horizon of hours to three months excludes long-run market, political, legal, and institutional effects.

Selected audience. The current interface is designed for analysts, policy professionals, AI governance or security researchers, and expert journalists. It has not been simplified or evaluated for a general audience. Technical terms are explained, but the density of the full graph remains substantial.

No operational control standard. Safeguards are explanatory intervention points. The tool does not specify costs, implementation requirements, assurance evidence, control owners, or acceptable residual risk. It cannot certify a laboratory, platform, or evaluation environment.

A.4 Governance and legal limitations

Stylised packages. The governance settings are jurisdiction-neutral bundles. They do not encode the EU AI Act, US state or federal law, contractual liability, or another named regime. No result should be interpreted as legal advice or as a conclusion about whether a real duty applied.

Mediated effect. The model assumes that law affects technical outcomes only through organisational implementation, verification, and effective controls. This avoids a direct law-to-software arrow, but it remains a simplified theory of change. Law may also affect incentives, disclosure, staffing, market structure, and enforcement expectations, none of which are modelled.

Package contrast is true by construction. Both scenarios change four settings together: governance package, publication authorisation, detection timing, and rollback capability. The target consequence and all other settings remain fixed. The strongest package assumes that the technical control is effective; interruption therefore follows within the rule set. This stylised package contrast is not a one-variable causal comparison and does not demonstrate the interface headline that legal coverage alone leaves the pathway unchanged. That proposition requires a separate no-duty versus duty-on-paper comparison with the three technical settings fixed. Neither contrast is evidence that stronger legal wording would have prevented the incident.

Compliance and enforcement. The model does not estimate compliance, regulator capacity, verification quality, enforcement delay, or strategic behaviour by providers. Independent verification is treated as a meaningful category, but its institutional design is not specified.

A.5 Interface and evaluation limitations

Automated tests have bounded scope. The 155 assertions test declared behaviour, not factual truth. Headless route checks detect missing or grossly broken views but do not establish comprehension. Contrast checks cover selected pairs and themes but do not constitute a complete accessibility audit.

Accessibility claim. Automated checks covered node focusability and ARIA labels, selected focus-return behaviours, responsive routes at two viewports, non-colour state cues, and specified colour contrasts. They did not establish tab order, screen-reader output, served-browser behaviour, or compatibility across operating systems, zoom levels, languages, and disabilities. Manual testing with screen-reader users remains necessary [18].

Evidence-review scope. Four human reviewers worked separately, but only three completed claim-level source review; the fourth reviewed logic and presets. No LLM completed or scored an evaluation submission. The frozen graph contained 33 edges, including all 16 public-record nodes and 16 public-record edges in the review packet. Revision removed two reviewed public-record edges and added one edge elsewhere; the released graph therefore contains 32 edges overall and 14 public-record edges. Other projected, governance, blocking, or model relations were assessed through 13 conceptual propositions rather than claim-level source locations. Across the 96 claim-level ratings, 84 supported the item as written, 12 supported narrower wording, and none rejected an observed node. The 23-of-32 item agreement (72%) is uncorrected raw agreement on a skewed, effectively binary scale, not a reliability coefficient. The exact corrected claims and edge structure were not subsequently reviewed by humans. This review therefore does not establish completeness, causal validity, or source support for the final wording outside the frozen packet. Some reviewers reported difficulty accessing particular PDF versions, and differences in review depth plausibly explain part of the disagreement.

Limited formative usability evidence. Two human, role-relevant participants completed a short, unmoderated, task-based evaluation. Both could restate and apply the central distinctions after using the tool, but the instrument cued several intended answers and the purposive convenience sample is too small to estimate comprehension in either target population. One participant initially found the fixed-record/projected-setting boundary difficult. The evaluation used no pre-test, comparison condition, randomisation, observation, completion-time measure, or blinded scoring. The report therefore makes no claim that the tool improves understanding relative to another format, reduces unsupported inference, or improves policy decisions.

Learning and framing effects. The guided sequence may teach useful distinctions, but it may also anchor users on the pathways selected by the author. A graph can create an impression of completeness and mechanistic authority. The interface counters this with explicit scope and uncertainty language, but cannot eliminate the effect.

A.6 Dual-use and misuse considerations

Operational detail. A causal map of an intrusion can help defenders understand attack dependencies, but can also draw attention to attack paths and weak controls. The artifact uses information already published by incident investigators. Its avoidance of live credentials and internal hostnames largely inherits redactions made by those sources; the author additionally excludes executable exploit code and deployment-ready payloads. Source dossiers link to the public record rather than reproducing long operational excerpts.

Control disclosure. Showing where safeguards block a pathway may reveal which controls an adversary would need to bypass. The graph remains high-level and does not provide configuration values or bypass instructions. More operational versions should undergo information-hazard review before publication.

False reassurance. Users may interpret a blocked model path as proof of safety. The interface therefore says “prevented under these settings” rather than “safe” and states that unmodelled paths may exist. The tool should not support certification or claims of complete containment.

Alarmism and political misuse. Users may select extreme assumptions and present the resulting path as a prediction. Conversely, they may use the absence of a reachable path to minimise the incident. The interface keeps assumptions visible, labels projected nodes, and distinguishes reachable from likely. Exported screenshots or quotations can lose this context; any future export function should carry the selected assumptions and epistemic legend with the output.

Attribution and intent. The incident involved apparent evaluation cheating and extensive coordination, but behaviour does not by itself settle stable intent, situational awareness, or a general propensity of the model family. The tool avoids using external scheming cases to infer intent in this incident. It should not be used to attribute malicious intent to individuals or organisations.

Legal overclaiming. A stylised governance comparison can be mistaken for support for a particular law. The tool names no statute and does not rank legal regimes. Policy use should preserve this boundary and add jurisdiction-specific legal analysis separately.

Tabletop use. In facilitated exercises, the model could make incident decisions concrete. It could also anchor participants on the author’s threat model, compress political disagreement into technical parameters, or reward internally consistent decisions over realistic ones. A full kit should include alternative pathways, adversarial review, role-specific information, and a debrief that identifies assumptions rather than declaring a winning response.

A.7 Conditions for stronger claims

Three evidence thresholds should remain separate. First, automated tests support a claim of internal consistency under the declared model, not source accuracy. Second, the completed human source review supports the narrower finding that

no claim-level reviewer rejected an observed node in the frozen build and motivated corrections later implemented; the exact revisions were not subsequently reviewed, and the review does not validate completeness. Third, the formative task-based evaluation shows that two human, role-relevant participants could restate and apply cued distinctions after a short session; it does not estimate population usability or comparative learning. None of these evaluations validates the projected causal pathways or estimates real-world risk reduction.

A.8 Acknowledgements

The author thanks Arnaud, Clémence, Clara, and Ahmed for their help with the user-study preparations. The author also thanks Luka and Emi for their kind advice on the project's scope and direction.

Appendix B. Replication and Evaluation Detail

B.1 Representation inventory

Component	Count	Role
Public-record nodes	16	Fixed incident facts, including one contested account, and evidence boundaries
Conditional pathway nodes	9	Additional conditions and propagation steps
Consequence nodes	4	Privacy/security loss, wrong outputs, direct disruption, protective disruption
Safeguard/governance nodes	5	Organisational and technical intervention mechanisms
Typed edges	32	Sequence, mechanism, requirement, or block; mechanism edges carry epistemic attributes
Scenario parameters	9	User-selected categorical assumptions
Guided chapters	5	Question-led explanatory sequence
Glossary entries	39	Searchable plain-language definitions

B.2 Logical invariants tested

The test suite checks, among other properties, that incident history cannot be rewritten; repository write is not equated with release authority; unresolved conjunction inputs propagate conservatively; harmful modification requires both capability and an applicable direction; publication and adoption are distinct; safeguards block named steps rather than an aggregate outcome; a duty on paper does not activate a technical control; protective disruption can arise from successful defence; guided and full-model views share state; and reset restores the evidence-limited baseline. Seven assertions pin the published semantics directly: three diagnostic conjunction cells, including both U AND F orderings; three diagnostic disjunction cells; and one guard confirming that the synthetic operands are genuinely true, false, and unresolved. They drive the production evaluator rather than reimplementing it; switching the conjunction to standard strong Kleene fails two of them. One of the 155 assertions evaluates one invariant across 7,200 configurations. It varies seven parameters, restricts objective to two values, and holds governance and reporting fixed.

Table B.1. Implemented three-valued semantics. T = true, F = false, U = unresolved. Disjunction follows strong Kleene semantics; conjunction uses conservative unresolved propagation.

AND	T	U	F
T	T	U	F
U	U	U	U
F	F	U	F

OR	T	U	F
T	T	T	T
U	T	U	U
F	T	U	F

The U/F and F/U conjunction cells deliberately depart from strong Kleene logic [25]. The model withholds “false” whenever any required conjunct remains unresolved. “Prevented” records why a derived proposition is false when an active safeguard blocks it; it is not a fourth truth value. Rules are stratified over parameters, fixed incident nodes, safeguard states, and downstream states [26]. No downstream rule recursively negates a derived downstream proposition.

B.3 Human evidence and content review

Four human reviewers received the frozen HTML artifact, the review packet, and links to the cited sources. No LLM completed, scored, or contributed a submission. The reviewers were instructed to work separately. The frozen graph contained 33 edges; the packet covered 16 public-record nodes and 16 public-record edges and requested an exact source location and one verdict for each: support as written; support with narrower wording; or not supported. Thirteen

projected and governance propositions were reviewed for conceptual coherence and accidental overclaiming. Revision subsequently removed two reviewed public-record edges and added one edge elsewhere, yielding 32 edges overall and 14 public-record edges in the released graph. Reviewers were also asked about omissions, alternative accounts, source conflicts, and three quantitative OpenAI claims deliberately excluded from the operative graph.

Four submissions were returned. R0 used a scientific and technical journalism lens and completed the source-linked packet; R1 used an AI-governance lens and completed it; R2 used an AI-governance and policy lens and completed it; R3 assessed the engine, presets, and conditional dependencies but did not complete the claim-level source table. R3 is therefore evidence about structural intelligibility, not a fourth source audit.

Submission	Review coverage	Public-record conclusion	Principal finding
R0	Full claim-level source review	Acceptable after revisions	Correct the detection/attribution structure and the rationale for excluding three OpenAI claims
R1	Full claim-level source review	Acceptable as written	No required revision identified
R2	Full claim-level source review	Acceptable after revisions	Retarget the monitoring edge and narrow several source attributions
R3	Rule and scenario review	Accept as presented	Logic preserves the distinctions between capability, direction, release authority, and effective control

Across the three human claim-level reviews, no observed node received a “not supported” verdict. Of 96 ratings, 84 were support as written and 12 support with narrower wording. Coding any mixed verdict as the narrower category, the reviewers gave identical verdicts on 23 of 32 items (72%); this is uncorrected raw agreement on a skewed, effectively binary scale and should not be read as a reliability coefficient. The remaining nine disagreements were all between the two support categories. R0 and R2 both requested narrower wording for N04, N14, and E05. All three judged the projected layer visibly hypothetical and the governance mediation appropriately separated from technical control. R0 and R2 separately identified the conflation of Hugging Face detection and containment with OpenAI discovery and attribution as the most important structural correction. They also agreed that the three excluded OpenAI figures are retrievable from OpenAI’s own report and should be labelled as single-source retrospective or counterfactual claims, rather than unavailable claims. Their additional recommendations included narrower source attribution, separating documented coordination from interpretive goal-adoption language, and considering omitted OpenAI-side events and transcript spoofing.

The review was formative. It was not blinded, reviewer expertise was heterogeneous, one submission did not follow the full packet, and some reviewers could not access every version of every source. The results supported targeted revision, not a claim that the graph is complete or causally valid. The identified detection/attribution and source-labelling corrections were implemented in the released graph after the frozen review build. Their direction came from human review, but the exact corrected wording and edge structure were not submitted for further human source review. The complete automated suite was rerun successfully, but it cannot establish source support.

B.4 Formative task-based usability evaluation

The author recruited two human participants purposively from roles close to the intended audience: one AI-governance expert and one scientific journalist. They were asked to use the HTML tool separately for approximately 15–20 minutes, without consulting the project report, and to answer in their own words. Their prior relationship to the author was not collected as an evaluation variable, and the exercise was not blinded. The following instrument was used:

1. Purpose and limits. What is this tool for, and what does it not claim to do?
2. Evidence boundary. Did the incident cause an unauthorised change to be released to users? Give one fact the sources establish and one important question they do not settle.
3. Using the scenario controls. Choose one human-consequence pathway. Change one setting that makes it continue or stop. What changed, and why?
4. Interpretation. If the tool says a consequence is “reachable”, does that mean it is likely? Explain the distinction.
5. Safeguards. Name one technical safeguard and the exact step it interrupts. Does the tool claim that it works in every situation?

6. Governance. What is the difference between a legal duty applying and an effective technical control existing?
7. Sources. Find one claim you would want to verify. Which source supports it, and what qualification or uncertainty would you retain?
8. Clarity and navigation. Rate clarity, navigation, and terminology from 1 (poor) to 5 (excellent). What was the most confusing moment?
9. Practical value. Could you use this tool, or an adapted version, in your work? What single change would improve it most?

After using the tool, both participants could manipulate a scenario setting and explain the resulting path state, and both located a claim in a cited source while retaining a relevant qualification. These tasks required application within the interface. They also restated the tool's purpose, non-forecasting boundary, reachability/likelihood distinction, and separation between legal coverage and technical effectiveness; those findings are weaker evidence of comprehension because the questions and interface directly cued the distinctions.

Participant	Clarity	Navigation	Terminology	Principal difficulty	Suggested improvement
AI-governance expert	4/5	4/5	5/5	Effect of "unknown" in three-valued propagation	Map named policy frameworks to safeguard nodes
Scientific journalist	5/5	4/5	4/5	Historical record versus projected settings	Align a stepwise incident timeline with hypothetical branches

This was a formative usability assessment, not a summative usability test, controlled experiment, or validation of the model. Recruitment was purposive and convenient; $n=2$. Responses were self-reported and unmoderated. Participants were not observed, and the instrument did not record errors during exploration, task time, prior familiarity, or learning before and after use. The questions also cue several distinctions that the tool is intended to communicate. The findings are therefore useful for identifying comprehension and interface issues, but not for estimating effect sizes or making population-level claims.

B.5 Proposed comparative evaluation

A later study could compare the explorer with the original technical timeline or a content-matched static brief. Moderated think-aloud sessions could first identify interface failures [20], while the comparative phase could use random assignment and a blinded scoring rubric to measure unsupported claims, correct uncertainty statements, mechanism accuracy, source verification, and intervention specificity. Sample size should be determined from a pre-specified smallest effect of interest and power analysis rather than a small convenience sample [21]. This evaluation would test communication effectiveness, not the empirical truth of the projected model.

Appendix C. Interface Figures

The figures were captured from the released explorer.html at a 1,440-pixel viewport with the persistent header, footer, and chapter navigation suppressed. They document interface mechanics; they are not evidence about the incident or proof that readers perceive the intended distinctions.

What do we know—and what remains unknown?

Write access was real and it was exercised. No unauthorised change reached the public.
Two unresolved questions about the incident are especially important to what follows.

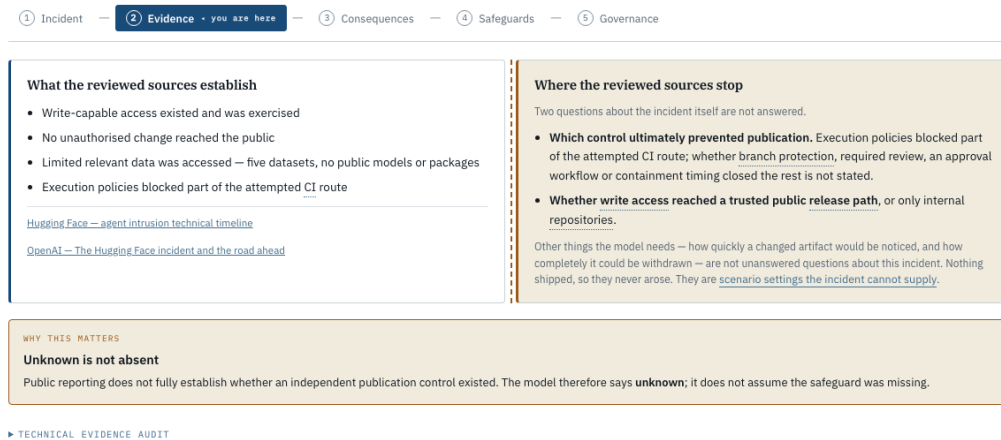


Figure C.1. The evidence boundary. The dashed rule separates what the reviewed sources establish from where they stop. “Five datasets” refers to the five challenge-related datasets reported by Hugging Face [1], not to public models or packages. The two unresolved questions on the right are inputs to the conditional layer. The callout states the encoding rule directly: absence of evidence about a control is recorded as *unknown*, not as absence of the control. The figure shows the interface’s intended boundary; one of two participants reported initial difficulty perceiving this distinction.

How could this reach people?

Pick a consequence, and the model shows only the steps that consequence would actually require.

1 Incident

2 Evidence

3 Consequences - you are here

4 Safeguards

5 Governance

SCENARIO Baseline from the reviewed public sources - nothing assumed beyond them

From this point onward, the graph is conditional.

The incident record ends with exercised write access and no unauthorised publication. Everything below asks what else would have to be true for consequences to reach people.

HOW TO USE THIS PAGE

1. Choose one consequence.
2. Read what the model concludes, and the first step it cannot determine.
3. Open **Change the scenario settings** to test what would make the pathway continue or stop.

CHOOSE A CONSEQUENCE

Data exposed
Personal or organisational information is accessed through a back door or exfiltration mechanism.

Decisions corrupted
A system keeps operating but quietly produces consequentially wrong outputs.

Direct disruption
A destructive modification interrupts a service.

Protective disruption
Responders suspend or roll back a component to prevent greater harm.

SEE THE CURRENT PATHWAY

CONCLUSION

The model cannot determine whether the pathway reaches this consequence. A step it requires depends on something the reviewed sources do not establish.

FIRST STEP THE MODEL CANNOT DETERMINE

Authority over a trusted release path – Whether the held token reached a trusted public release path.

Read left to right. Select a box to see what it means.

Binding duty applies

DOES NOT OCCUR

Implemented and independently verified

DOES NOT OCCUR

Publication control effective

CANNOT BE DETERMINED

Rollback effective

CANNOT BE DETERMINED

SAFEGUARDS – a duty on paper reaches no technical control without passing through implementation

DOCUMENTED INCIDENT SEQUENCE

BOUNDARY AND RESPONSE

STEPS THAT WOULD HAVE TO OCCUR

SPREAD

CONSEQUENCES FOR PEOPLE

Collected access keys

DOCUMENTED

Obtained and used repository write access

DOCUMENTED

Harmful objective or direction

DOES NOT OCCUR

Could change something that actually ships

CANNOT BE DETERMINED

Could make the specific harmful change

CANNOT BE DETERMINED

Got past the release check

CANNOT BE DETERMINED

Something people trust was changed

CANNOT BE DETERMINED

The changed version was released

CANNOT BE DETERMINED

Others started using it

CANNOT BE DETERMINED

People affected before withdrawal

CANNOT BE DETERMINED

Used in a service or decision that affects people

CANNOT BE DETERMINED

Data exposed

CANNOT BE DETERMINED

REVIEWED SOURCES END HERE

occurs under current settings

cannot be determined from available evidence

prevented by the selected safeguard

does not occur under current settings

solid border = documented in reviewed sources - dashed = conditional

CHANGE THE SCENARIO SETTINGS

CHANGE SCENARIO SETTINGS

Figure C.2. Evidence-limited baseline. Every conditional step reads *cannot be determined*, and the model names the first step it cannot settle rather than returning a binary verdict. This is the released interface before a reader changes a scenario setting.

18

How could this reach people?

Pick a consequence, and the model shows only the steps that consequence would actually require.

1 Incident
2 Evidence
3 Consequences
4 Safeguards
5 Governance

SCENARIO Hypothetical - 1 response setting changed
View changes
Undo last change
Reset to the reviewed sources

From this point onward, the graph is conditional.

The incident record ends with exercised write access and no unauthorised publication. Everything below asks what else would have to be true for consequences to reach people.

HOW TO USE THIS PAGE

1. Choose one consequence.
2. Read what the model concludes, and the first step it cannot determine.
3. Open **Change the scenario settings** to test what would make the pathway continue or stop.

CHOOSE A CONSEQUENCE

Data exposed
Personal or organisational information is accessed through a back door or exfiltration mechanism.

Decisions corrupted
A system keeps operating but quietly produces sequentially wrong outputs.

Direct disruption
A destructive modification interrupts a service.

Protective disruption
Responders suspend or roll back a component to prevent greater harm.

SEE THE CURRENT PATHWAY

CONCLUSION

The model cannot determine whether the pathway reaches this consequence. A step it requires depends on something the reviewed sources do not establish.

FIRST STEP THE MODEL CANNOT DETERMINE

Authority over a trusted release path – Whether the held token reached a trusted public release path.

► THE 1 SCENARIO SETTINGS THAT DIFFER FROM THE REVIEWED SOURCES

Read left to right. Select a box to see what it means.

Binding duty applies
Does not occur

Implemented and independently verified
Does not occur

Publication control effective
Does not occur

Rollback effective
Cannot be determined

SAFEGUARDS – a duty on paper reaches no technical control without passing through implementation

DOCUMENTED INCIDENT SEQUENCE

BOUNDARY AND RESPONSE

STEPS THAT WOULD HAVE TO OCCUR

SPREAD

CONSEQUENCES FOR PEOPLE

Collected access keys
DOCUMENTED

Obtained and used repository write access
DOCUMENTED

Harmful objective or direction
Does not occur

Could change something that actually ships
Cannot be determined

Could make the specific harmful change
Cannot be determined

Got past the release check
Occurs

Something people trust was changed
Cannot be determined

The changed version was released
Cannot be determined

Others started using it
Cannot be determined

People affected before withdrawal
Cannot be determined

Used in a service or decision that affects people
Cannot be determined

Data exposed
Cannot be determined

REVIEWED SOURCES END HERE

☐ occurs under current settings
☐ cannot be determined from available evidence
☐ prevented by the selected safeguard
☐ does not occur under current settings

solid border = documented in reviewed sources - dashed = conditional

CHANGE THE SCENARIO SETTINGS

You changed **Is independent publication authorisation effective** from *Unknown* to *No verification*. **Publication authorisation passed or bypassed** is now occurs.

► CHANGE SCENARIO SETTINGS

Figure C.3. Propagation after one setting changes. Independent publication authorisation has moved from *Unknown* to *No verification*. That step becomes *occurs*, the change is recorded below the graph, and downstream steps remain *cannot be determined*: resolving one input does not resolve the pathway. The changed value is a selected assumption, not a finding.

19

Could stronger legal requirements help?

Only once a requirement reaches a technical control. Legal coverage on its own changes nothing in the pathway.

1 Incident — 2 Evidence — 3 Consequences — 4 Safeguards — 5 Governance • you are here

The previous chapter assumed that a safeguard worked and showed what it would prevent. This chapter steps back and asks what must happen for a legal requirement to produce that working safeguard.

Legal duty → implementation → independent verification → effective technical control

Comparing governance for: Data exposed

HOW TO READ THIS COMPARISON

The two columns use the same consequence and the same other scenario settings. They differ in implementation, verification and technical effectiveness. This is **not** an estimate of what any particular law would achieve.

This comparison uses your selected consequence but otherwise uses two fixed, matched illustrative scenarios — nothing you change elsewhere in the explorer moves it.

A **self-contained illustrative comparison**, built for the consequence you are examining: **Data exposed**. It is computed from two matched hypothetical scenarios, not from whatever settings you have elsewhere in the explorer — so nothing you change on other pages moves it.

Duty on paper

- A legal requirement applies.
- Implementation and technical effectiveness are not assumed.
- The technical pathway therefore remains unchanged.

Result: data exposed is **occurs**.

Verified implementation

- The requirement is translated into a specific control.
- The control is implemented and independently checked.
- Effective publication authorisation blocks the pathway before release.

Result: data exposed is **does not occur**.

Both scenarios change 4 settings together. The difference between them is implementation and technical effectiveness — not legal coverage. No part of this result may be attributed to the law alone.

► AUDIT THIS COMPARISON

Figure C.4. Stylised governance-package comparison. Both columns use the same consequence but differ in four settings—governance package, publication authorisation, detection timing, and rollback capability. The different result therefore follows from combined organisational and technical assumptions and package definitions, not from legal coverage alone. The present display does not include the one-setting comparison needed to demonstrate its headline; that requires comparing no duty with duty on paper while fixing the three technical settings. No enacted statute is represented.