



QENDROVA
Decision evidence layer

QENDROVA An Executable Decision Boundary Protocol for AI Incident Evidence

Flamur Koca

QENDROVA (independent project)

With Apart Research

Abstract

Incident response can fail when a fact, inference, absence claim, or control proposal acquires more decision authority than its evidence permits. QENDROVA makes that boundary executable. In the public OpenAI-Hugging Face agent intrusion, it separates 20 atomic claims across five decision rooms, preserves source scope and unresolved remainder, and requires human confirmation, rejection, or hold before a final review state. Four resolvable questions bind an input, a tomorrow-runnable check, a resolution condition, a HOLD boundary, and a falsifiable prediction. A local dossier retains proposal, human action, revision, rationale, support scope, source binding, and integrity fingerprint; later evidence enters append-only. Deterministic implementation suites pass 28/28, 15/15, 12/12, 18/18, and 714/714 cases. These counts establish protocol behavior, not incident truth. QENDROVA's contribution is selective consequence control: supported evidence remains actionable while unsupported closure cannot borrow authority.

1. Introduction

Public accounts by Hugging Face and OpenAI describe an AI system crossing technical and organizational boundaries during a model evaluation. The response problem is not only reconstructing reported events. It is deciding how far each piece of evidence may change an operational state. A source-bounded observation can become a global conclusion; a plausible mechanism can become a claimed cause; 'not identified' can become 'did not occur'; or a technical proposal can acquire authority that remains human or institutional. These transitions can produce both overreaction and false closure. Source scope: Hugging Face technical timeline and OpenAI incident materials linked in Section 2.

QENDROVA addresses this decision-boundary problem. It represents incident statements as separate claims, binds each claim to its source and scope, produces a bounded status proposal, and requires an explicit human action before a final review state is recorded. The contribution is operational: an incident lead can use the result to identify what requires review, what remains source-bounded, what stays unresolved, and which decisions the system is not authorized to make.

Our main contributions are:

- an executable 20-claim review protocol across five decision rooms, preserving both evidential reach and non-reach;
- four immediately runnable checks and four falsifiable predictions with explicit disconfirming outcomes;
- a locally re-verifiable v2 dossier that preserves claim text, proposed effect, source binding and scope, system proposal, canonical rationale, human action, final status, register revision, and fingerprint;
- an append-only update gate that admits later official evidence claim by claim without rewriting the frozen base register.

2. Related Work

The public record consists primarily of organizational incident accounts. Hugging Face published a technical timeline from the receiving platform's perspective, while OpenAI published the evaluation origin and later a broader technical report. These accounts overlap but do not have identical visibility or independence. Support is therefore recorded only inside the named source scope; source count alone never establishes independent corroboration or forensic ground truth. The Apart Track 2 brief asks participants to reconstruct what the incident establishes and does not establish, then produce resolvable questions, checks usable tomorrow, and causal explanations that predict something.

Conventional narrative summaries are useful for communication, but they can hide the transition from observation to inference and make revisions difficult to audit. QENDROVA's narrower contribution is an executable decision layer: it does not replace incident-response platforms, source forensics, or expert judgment. It controls when evidence is allowed to change a claim and preserves an explicit remainder when the available record cannot close the question.

Primary-source incident record: [Hugging Face technical timeline](#) · [OpenAI incident announcement](#) · [OpenAI later report](#)

2.1 Incident film: established inside scope / not established

- Established inside named source scope: an evaluation-origin account, a receiving-platform timeline, reported technical paths and observed effects, and later control proposals. These entries remain attributable to the source that reports them.
- Not established by the public record alone: complete forensic ground truth, source independence, malicious intent, a globally closed impact denominator, or the operational sufficiency of announced remediation.

Decision consequence: supported observations may trigger bounded review now; stronger causal, absence, completion, or authority claims remain REVIEW, HOLD, or UNKNOWN until their resolver is satisfied.

3. Methods

3.1 Claim register and status vocabulary

We decomposed the public record into 20 atomic claims and placed them in five rooms: evaluation context; technical path and reach; limits of the forensic record; intent, cause, and counterfactuals; and controls, future implications, and authority. Each entry stores the claim, kind, named source IDs, proposed status, proposed effect, and any claim-specific support note. The UI displays the claim, source chips, proposal status, canonical rationale, and claim-specific note before a reviewer acts.

Status	Operational meaning
SUPPORTED	Supported inside the named source scope; source count alone is not independent corroboration.
SOURCE_BOUNDED	Useful inside one source's visibility; not promoted beyond that scope.
REVIEW	A plausible inference or control consequence that requires human examination.
HOLD	The proposed conclusion exceeds current evidence, scope, or authority.
UNKNOWN	The available evidence does not resolve the question.

3.2 Human authority and local dossier

For every claim the reviewer chooses confirm, reject, or hold. The original system proposal remains visible even when the reviewer changes it. The Windows package runs on local loopback, verifies its files before launch, uses a fresh temporary browser profile, and exports a local JSON dossier. The v2 dossier stores claim text, proposed effect, canonical rationale, support scope, source metadata, register revision, human action, and final status. A separate verifier checks those fields against the trusted register and rejects stale-hash tampering, re-hashed false proposals, invented rationales, and false revisions.

3.3 Questions, checks, and falsifiable predictions

Four resolution contracts cover transitive egress, cross-run coordination, signal to pause, and the denominator of negative claims. Each binds the required input, a tomorrow-runnable action, a bounded resolution condition, the condition that forces HOLD, and a falsifiable prediction. Four checks turn those questions into work

packages with an owner, evidence requirement, deadline, and explicit unresolved remainder. The four falsifiable predictions state observations that would count against the proposed mechanisms; they do not by themselves establish causal identification.

Resolvable question	Check someone can run tomorrow	Input · resolution · HOLD · prediction
Transitive egress	Send harmless canaries through every declared support route in an isolated environment.	Input: route map + retained logs. Resolve on path-matched evidence; otherwise HOLD. Predict: an undeclared proxy creates an unmatched hop.
Cross-run coordination	Run fresh agents against synthetic shared state and measure whether a marked artifact changes later action.	Input: fresh runs + synthetic shared state. Resolve on a bound read/write/action trace; otherwise HOLD. Predict: removing persistence removes the effect.
Signal to pause	Replay harmless severe signals and time correlation, paging, pause, and human-authorized restart.	Input: severe harmless signals + timestamps. Resolve on predeclared pause/restart evidence; otherwise HOLD. Predict: uncorrelated alerts miss the pause SLA.
Negative-claim denominator	Audit every public exclusion claim for asset set, time window, telemetry coverage, and excluded paths.	Input: exclusion claim + asset/time/telemetry/path denominator. Resolve only when all are explicit and testable; otherwise HOLD. Predict: later scope expansion breaks the claim.

3.4 Later evidence without silent replacement

The base register is frozen by fingerprint. CAP-085 accepts a later source only as an append-only update proposal. It records the old register fingerprint, new source, claim-level relation, previous and proposed status, rationale, and human decision. Publisher count is recorded but does not prove independence; explicit source dependency can keep a two-publisher claim SOURCE_BOUNDED. HOLD or UNKNOWN cannot become SUPPORTED without an explicitly satisfied resolver. Unaffected claims must retain their fingerprints.

3.5 Provenance of the project

Time boundary	Auditable work status
Pre-existing QENDROVA	Decision operators, local evidence infrastructure, authority separation, scope and UNKNOWN preservation.
Pre-sprint adaptation · 25 Aug–10 Sep 2026	Public-source extraction, 20-claim register, five-room review, questions/checks, verifier, update gate, Windows package, and pre-sprint report content.
Sprint window · 11 Sep 2026	Append-only integrity work: comparison Contract 009; public-source snapshot Manifest 001; live form and metadata audit; report revision. Human baseline run remains NOT_RUN.



Substantial incident-adaptation content existed before the sprint. During the sprint window, integrity and submission work was recorded append-only: comparison Contract 009, a hash-bound two-source snapshot denominator, identity-corrected pre-run human artifacts, live submission-form and metadata audits, the scrubbed v1.07 package, the local validation receipt, the preserved rejection chain for Candidates 010 through 015, and this Candidate 016 report revision. Contract 009 remains PRECOMMITTED_NOT_RUN; no human baseline result exists. Template use indicates preparation for the Apart Research sprint and does not imply endorsement.

Figure 1. The system proposes a bounded status; human review authorizes the final dossier state. Later sources enter through a visible revision path.

4. Results

The revised base incident register contains three SUPPORTED claims, six SOURCE_BOUNDED claims, five REVIEW obligations, three HOLD claims, and three UNKNOWN claims. The signature result is dual: Hugging Face's reported five-dataset boundary remains useful as SOURCE_BOUNDED, while the stronger inference that no wider impact occurred remains HOLD. QENDROVA therefore preserves a positive statement and justified non-closure at the same time.

Deterministic implementation suite	Passed	What the fixture establishes
Incident review and package	28/28	Claim completeness, five-room structure, local runtime, source and authority boundaries.
Dossier verifier	15/15	Valid dossiers pass; stale, re-hashed, rationale, revision, and missing-note tampering fail against trusted contracts.
Resolution contract	12/12	Questions bind scope, window, evidence, metric, threshold, result vocabulary, and authority.
Selective evidence update	18/18	Frozen base, claim-level revision, human action, publisher independence, and untouched fingerprints.
Full QENDROVA regression	714/714	Fifty deterministic suites retain the wider operator behavior with zero failed fixture cases.

The 26 August OpenAI report was then treated as a later source rather than silently merged into the July base. In the update fixture it can add a second organizational perspective to IC-06 and IC-07 after human confirmation. IC-14 remains REVIEW because a stronger narrative does not itself prove the causal interpretation. IC-18 remains UNKNOWN because announced controls do not prove operational sufficiency. Claims outside the update keep their original fingerprints.

These are software and protocol results, not a claim that all incident facts are scientifically verified. They establish repeatable behavior in the bound artifact: the system preserves proposal/decision separation, resists several integrity failures, and prevents later evidence from acquiring global authority merely because it is newer or more detailed.

5. Discussion and Limitations

QENDROVA's practical value is selective consequence management. Incident teams need action before every uncertainty is resolved, but acting does not require converting every uncertainty into certainty. The protocol allows a supported event to trigger a bounded review while preserving independent information and unresolved causes. That can reduce both false preservation and indiscriminate revalidation, provided the claim decomposition and source binding are competent.

Limitations

The 20-claim base is human-curated from public organizational accounts and contains no private logs. SUPPORTED means support inside the named source scope; source count alone does not establish independence or forensic truth. Technical test counts measure deterministic software invariants and fixture paths, not scientific accuracy, statistical superiority, reviewer expertise, control effectiveness, or legal responsibility. No empirical baseline comparison has yet been run. The current package has been rebuilt and locally re-verified but lacks a documented second-device run and independent subject-matter review. The interface demonstrates a bound incident workflow, not automatic extraction from arbitrary reports. SHA-256 protects content integrity when a trusted fingerprint is available; it does not prove when the content existed.

Future Work

The next evidence should be a separate subject-matter review of all 20 bindings, a second-device and second-reviewer run with correction rate and setup time, and the precommitted comparison against a conventional summary or spreadsheet on unsupported closure, source-merging error, and reviewer time. Future incident cases should add independent time evidence for criteria, dataset identities, thresholds, and hashes before outcomes exist.

6. Conclusion

QENDROVA turns the boundary between incident evidence and decision into an executable artifact. In the public OpenAI-Hugging Face case it identifies what is supported inside named source scope, what remains source-bounded, what requires review, what stays unresolved, and which decisions remain outside system authority. Visible limits do not reduce the usable evidence: supported consequences remain actionable, while unsupported conclusions cannot borrow the same evidential weight.

Code and Data

The juror application is

QENDROVA-INCIDENT-DESKTOP-v1.07-WINDOWS-X64-20260911054000.zip (SHA-256 9b0219ea7143e5cba576634f2580d1cf9aeefb20e7265d7a65ec9807aa2ebfae). Its PACKAGE-MANIFEST.json binds 40 payload files; the manifest is the 41st physical file. The ZIP supplies the local 20-claim review, JSON dossier export, and trusted-contract verification path, subject to Windows x64 and Microsoft Edge. Comparison Contract 009 has SHA-256 8e96b6121b30b6be4c20df5b763960681d58668a053c7668a478711d91b93c9c and remains PRECOMMITTED_NOT_RUN. Validation Receipt 016 has SHA-256 86c42c86e736c0a761372d7158b3489191b5d5e7ae0774fd71d380078d1282c8 and binds the exact application, local deterministic results, Contract-009 integrity, and the hash-verified two-source baseline denominator while expressly excluding incident truth and empirical superiority. The bundle includes the append-only correction chain that rejects Candidates 010 through 015 where chronology, internal identity, validator-target, or source-denominator defects were found. Source snapshot Manifest 001 has SHA-256 6b4bce1667d80cd10bfd4472892ab7e86c4d525ab5319bfc38b5521d754bfd4f; the raw HTML bytes are retained for the controlled baseline but are not publicly distributed, so the public artifact alone cannot replay that byte comparison. A juror-accessible HTTPS delivery URL remains a final human HOLD. The package omits credentials, private operational endpoints, private logs, and runnable offensive chains.

Author Contributions

Flamur Koca conceived and directed QENDROVA as its solo founder, defined the decision operators and product requirements, selected the incident framing, challenged and revised claim boundaries, and reviewed the executable results. He is the sole human author and accepts responsibility for the submitted work. AI systems were used only as disclosed in the LLM Usage Statement.

References

- Hugging Face. 2026. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- OpenAI. 2026. OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- OpenAI. 2026. The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- OpenAI. 2026. Hugging Face Incident Technical Report. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- Apart Research. 2026. AI Incident Response Sprint. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>

Appendix A Limitations and Dual Use Considerations

Scope and validity. The artifact controls claim transitions inside a curated public case. It does not certify source completeness, truth, reviewer identity or competence, universal incident understanding, or legal authority. A human confirmation records a review action; it is not automatically substantive validation. The selective-update fixture demonstrates mechanics and does not represent production acceptance of the 26 August revisions.

Potential misuse. A curator could omit disconfirming sources, choose favorable claim boundaries, overstate publisher agreement, or use visible structure to create a false appearance of rigor. HOLD and UNKNOWN could also be abused to delay or suppress inconvenient supported claims; caution is not automatically neutrality. A reviewer could click through proposals without expertise. These risks are reduced by preserving original sources, proposals, corrections, holds, dissent, deadlines, and explicit resolution requirements, but they are not eliminated.

Information hazards. The public sources already contain operational detail. The QENDROVA package stores claim-level descriptions and defensive implications rather than exploit instructions, secrets, live indicators, or executable attack chains. Credentials, operational endpoints, unredacted identifiers, and unnecessary command detail must remain excluded from public packaging.

Safeguards. Package verification, local-only operation, explicit source links, bounded status vocabulary, human decision separation, append-only updates, and the trusted-contract dossier verifier make several classes of overreach or tampering visible. Independent review and time evidence remain separate requirements.

LLM Usage Statement

OpenAI Codex and ChatGPT, Google Gemini, and xAI Grok were used as development assistants for counterargument generation, source-checking prompts, implementation and debugging, test design, and editorial review. Flamur Koca defined the product and operator concepts, directed the work, challenged proposed interpretations, reviewed the resulting behavior, and accepts final responsibility for the work. Executed test results were produced locally; incident statements were checked against the cited public sources. AI output was never treated as evidence or final decision authority.