Human–Agent Attribution Discontinuity

Executive Overview

Version 0.5

Kevin V. Watson

August 25, 2026

Document information

Author: Kevin V. Watson

Version 0.5 DOI: 10.5281/zenodo.22099968

Prior deposited version: Version 0.4, DOI 10.5281/zenodo.21683276

Parent method: The Zemi Method

Canonical source: *Human–Agent Attribution Discontinuity: A Forensic and Assurance Methodology*, Version 0.5

Status: Public release. Not a validated standard, forensic doctrine, or legal attribution framework.

The problem HAAD addresses

AI agents act through accounts, credentials, tools, APIs, memories, retrieved content, sub-agents, and organizational control systems. When an action produces a consequential result, investigators and organizations often move too quickly from technical association to a conclusion about a person:

> *The agent acted through an account associated with Person P; therefore, P knowingly caused or authorized the action.*

That inference may be wrong. An authentication record may establish that a mechanism accepted account-linked factors without establishing who controlled the material interaction. A recorded approval may have been inherited from task state rather than independently captured. A broad delegation may establish permission for a category of activity without covering the recipient, amount, method, or timing of the action that occurred. Retrieved content or another agent may have supplied the effective instruction.

HAAD is a forensic and assurance method for determining how far the available evidence supports attribution of a consequential AI-agent action toward a natural person–and where that attribution must stop.

The central analytical object

HAAD examines **attributional transitions**. A transition is an evidentiary proposition connecting one node in the causal and attribution graph to another:

```
Natural person
→ device, session, or interaction
→ credential or authenticated identity
→ instruction or authorization
→ agent and configuration
→ memory, retrieval, delegation, and inter-agent influence
→ tool invocation
→ external action
→ consequence
```

The arrows are not assumptions. Each material arrow receives:

1. a precise proposition;
2. supporting and contrary evidence;
3. an articulated inference and warrant;
4. materially plausible competing origins;
5. a proposition-specific convergence analysis;

6. an adjudication status; and
7. an effect on the attribution boundary.

A **Human–Agent Attribution Discontinuity** exists where a material transition necessary to a proposed human-attribution conclusion is contradicted, assumed, or undetermined.

Protocol-mediated actions and MCP

Where an action is mediated by the Model Context Protocol, HAAD expands the technical path rather than treating "the agent used a tool" as a sufficient account:

```
Natural person → host application → model or agent → MCP client
→ MCP server → tool → downstream service → consequential action
```

Core MCP records may establish client authorization, tool discovery, transmitted arguments, server execution, or elicitation. Where adopted, the separately versioned MCP Tasks extension may add durable task-state records. Trace context, downstream receipts, and other observability records depend on the implementation or external systems. None of these records, without further evidence, establishes the natural person who controlled the material interaction, authored the instruction, understood the effective parameters, knowingly authorized the action, acted voluntarily, or intended the consequence.

Version 0.5 adds an MCP application profile without changing the fourteen canonical attribution dimensions. It requires preservation of effective tool definitions, package and handler versions, protocol messages, authorization and scope records, consent and confirmation events, parameter transformations, implementation trace context, adopted extension state, each downstream action, and the controls governing their integrity and completeness.

What HAAD produces

HAAD produces:

8. an attribution graph showing the material causal and evidentiary relationships;
9. proposition-level adjudications;
10. a discontinuity schedule identifying affected dimensions and failure conditions;
11. a bounded investigative-attribution conclusion;
12. a statement of the precise point at which human attribution stops; and
13. assurance and forensic-readiness requirements derived from the evidentiary gaps.

Stopping is a valid result. A technically established agent action need not become an attributed human action.

HAAD does not assign moral blame, civil liability, criminal liability, or legal responsibility. Its highest conclusion is investigative attribution. Legal and normative decisions remain with the competent decision-maker.

Core rules

No transitive attribution

Attribution does not automatically pass from an agent to a credential, from a credential to an account owner, or from an account owner to knowing authorization or intent. Every material transition requires its own evidence.

Narrow proof

Every artifact is credited only with what it proves. Authentication establishes accepted factors under the applicable mechanism. It does not alone prove natural-person control. A signature establishes use of a key under a trust model. It does not alone prove understanding, voluntariness, scope, or intent.

A subject-system statement proves that the statement occurred, not that the represented event, test, restoration state, explanation, or claimed reason is true. Recorded reasoning may support temporal and instruction-lineage analysis, but it does not self-prove faithful causal reasoning, awareness, belief, deception, or intent. A provider's confidence label proves that the provider reported an assessment; it does not become the practitioner's confidence without access to the supporting evidence and method.

An approval event proves that the recorded interface accepted a response. A cryptographic receipt proves the recorded technical binding under its trust model. Neither alone proves natural-person control, authorship, understanding, voluntariness, or knowing authorization of an undisclosed effective target or consequence. Incident-registry entries and technique mappings prove that a registry or catalog recorded an analysis; they do not independently prove inherited facts or create additional evidentiary streams.

Separate attribution dimensions

HAAD separates:

14. **AD-01 External action**;
15. **AD-02 Technical actor and execution**;
16. **AD-03 Credential or account**;
17. **AD-04 Human identity**;
18. **AD-05 Natural-person control**;
19. **AD-06 Authorship of the material interaction**;
20. **AD-07 Instruction origin**;
21. **AD-08 Knowing authorization**;
22. **AD-09 Voluntariness**;
23. **AD-10 Scope**;
24. **AD-11 Knowledge**;
25. **AD-12 Intent**;
26. **AD-13 Delegation**; and
27. **AD-14 Causal contribution**.

Establishing one does not establish the next.

These fourteen dimensions are the canonical register used by the taxonomy, workflow, conclusion layers, workpapers, pilot, and reference model. Individual and organizational investigative attribution are synthesized conclusions. Responsibility referral is a downstream output, not an attribution dimension.

Weakest-material-transition ceiling

The overall conclusion cannot be stronger than its weakest material transition. Strong downstream evidence cannot cure an upstream break necessary to connect that evidence to the proposed person. Evidence that an action was intentional, for example, does not establish whose intent it was when authorship of the material interaction or natural-person control remains unresolved.

The ceiling is path-specific. A weak transition from an operator cluster to an organization or state limits that higher-level conclusion without weakening a separately established technical-execution finding.

Proposition-specific independent convergence

Known attribution of natural-person control or knowing authorization ordinarily requires convergence among at least two proposition-relevant evidentiary streams with materially different capture or failure paths.

Independence is not a vendor count. Two logs from one provider may be independent for a proposition if they observed it through genuinely distinct mechanisms. Records from different providers may remain dependent if one merely copied the other's assertion.

A single-stream known-established exception requires a structured showing addressing directness, integrity, completeness, alteration capability, expected contradictory records, competing explanations, and residual failure risk. Enhanced application requires express second-practitioner concurrence.

Competing origins

Alternatives are not included merely because they are imaginable. They must pass:

28. a **plausibility gate**–technical feasibility plus a case-specific evidentiary or contextual anchor; and
29. a **materiality gate**–the alternative could change a material transition or attribution boundary.

Authorization and scope

An **Authorization discontinuity** exists where no valid grant covers the category of action.

A **Scope discontinuity** exists where a valid category-level grant exists but the action exceeds a recipient, amount, method, time, target, duration, tool, objective, or onward-delegation boundary.

The absence of an express prohibition does not establish authorization, and it does not by itself establish that the action was prohibited. A click, confirmation, continued session, or failure to intervene must be tested against what the person was shown and what effectively occurred. Permission to execute a command string does not establish knowing authorization of a materially different target, state basis, blast radius, or consequence.

Testimonial discipline

An authenticated interview proves that the recorded statement was made. It does not automatically prove the truth of every assertion. Admissions, denials, recollection failures, and exculpatory explanations are separately evaluated for interest, contemporaneity, specificity, consistency, and corroboration.

Operative state and later constraints

Attribution is evaluated against the instruction, permission, component, code, definition, policy, and configuration state operating at the material transition. A later freeze, revocation, or narrowed approval changes the recorded boundary only when its authority or authorship, delivery, timing, applicability, and continued operation are established. Continued technical capability does not establish continued authority.

Capability, authority, and enforcement

The ability to perform an action, the authority to perform it, and the control that should enforce the boundary are separate propositions. A failed control may support a readiness or organizational finding without establishing knowing human authorization of the historical action.

Multiple actions and implementation integrity

One invocation can produce an intended action and an unauthorized secondary action. Each consequence is adjudicated separately. A tool description or schema proves what was represented to the model or client, not that the effective package, handler, configuration, or downstream request conformed to it.

Attempt, execution, completion, effect, harm, mitigation, and recovery are also separate propositions. Recovery does not erase an established interruption or deletion, and an unsuccessful ultimate attack may still contain completed intermediate actions.

For command-mediated and infrastructure actions, HAAD reconstructs the effective target through the working directory, expansions, environment, state files, plans, credentials, account context, downstream requests, and affected resources. The state informing a decision is tested for provenance, version, freshness, completeness, and fitness.

Layered and aggregated attribution

Technical executor, account operator, natural person, analytic cluster, organization, and sponsor or state are separate paths. External confidence labels remain source-attributed when the supporting method or evidence is inaccessible.

Claims that AI performed a stated percentage of work must disclose their denominator, measurement method, coverage, exclusions, and uncertainty. Such a percentage does not allocate intent, authorization, causation, or responsibility. Human approvals at target selection, escalation, credential use, release, exfiltration, or other material gates are mapped separately.

Findings established for one event, target, session, or sampled episode are not generalized to a campaign without a stated aggregation rule and adequate evidence coverage.

Actions, runs, sessions, episodes, incidents, campaigns, and populations are counted separately. Repeated runs are not treated as independent when they share persistent accounts, artifacts, memory, infrastructure, instructions, operators, or another causal influence.

Provider, model family, model version, safeguard configuration, scaffold, agent instance, session, run, tool process, external account, and downstream executor are separate technical identities. Organizations are likewise decomposed into decision-makers, operators, control owners, monitors, responders, and other functions where their authority, knowledge, or contribution differs.

Two-axis discontinuity classification

HAAD separates:

30. **Affected attribution dimension:** what material proposition failed.
31. **Failure condition:** why continuity failed or could not be demonstrated.

A finding may therefore be expressed as:

> *Authorization discontinuity caused by dependent provenance and unresolved intervention.*

The primary discontinuity is the earliest material transition on the scoped causal path whose failure independently prevents the attribution dimension in issue. Other classifications are recorded as contributing, causal, or downstream.

Delegated Risk Tolerance

A Delegated Risk Tolerance record documents the categories, severity, limits, and conditions of risk knowingly accepted when a person delegates discretion to an agent.

It is evidence, not a waiver and not a conclusion. Its weight depends on specificity, intelligibility, actual authority, accurate capability disclosure, currency, affirmative acceptance, operational control linkage, and whether the resulting action remained within the stated boundaries.

A DRT does not transfer responsibility for system design, model changes, enforcement, logging, or organizational risk acceptance to an employee. In some cases, it may be stronger evidence of organizational notice than of individual responsibility.

Individual and organizational paths

Failure to establish individual attribution does not automatically establish organizational attribution. The organizational path is separately tested through propositions concerning:

```
Deployment decision
→ capability and authority design
→ organizational risk acceptance and notice
→ control ownership
→ monitoring and enforcement
→ agent operation
→ consequential action
```

The paths may be evaluated in parallel, but neither is a fallback assumption. The organization is not treated as one mind. Deployment, risk, operation, identity administration, control ownership, monitoring, and incident response are assigned to the functions the evidence supports.

Proportional application

HAAD supports three application levels:

32. **Triage:** provisional action, attribution path, evident discontinuities, competing origins, and boundary.
33. **Standard:** full graph, evidence and warrant analysis, competing-origin testing, adjudication, discontinuity schedule, and bounded conclusion.
34. **Enhanced:** all workpapers plus formal preservation, tool validation, expanded competing-origin testing, organizational analysis, and independent review.

Triage cannot support a final adverse finding about a person. Matters affecting rights, employment, reputation, access, legal position, or significant financial interests should escalate to Standard or Enhanced application.

Testing status

Version 0.5 is prepared for structured-scenario testing.

As of August 21, 2026, no independent-practitioner data exist for any HAAD version. Round 1 and Round 2 are planned and have not been conducted. Version 0.4 and Version 0.5 revisions are based on review and structured scenario analysis, not empirical pilot findings. The pilot evidence packs and sealed reference model are testing-design instruments, not completed results. The MCP application profile is based on review and specification analysis and remains subject to practitioner testing.

The public evidence packs and controlled reference companion are testing-design exemplars. Once a practitioner has seen them, they cannot serve as unseen Round 2 instruments. Public incidents, incident registries, technique mappings, and research corpora may seed new scenarios, but recognizable facts and published conclusions cannot become the hidden answer.

The Stage 1 pilot design evaluates:

35. agreement on ceiling-setting transitions;
36. agreement on the primary affected attribution dimension;
37. proposition-specific evidentiary-stream treatment;
38. agreement on the attribution boundary;
39. independent defensibility of the author-developed reference model; and
40. the ability to recognize both genuine discontinuities and a properly established attribution with no demonstrated discontinuity within scope;

41. the separation of technical, individual, organizational, and sponsorship ceilings;
42. the decomposition of multiple actions and subject-system representations;
43. tool-description versus implementation integrity;
44. quantitative activity-share and human decision-gate treatment; and
45. episode-level versus campaign-level aggregation;
46. approval and non-intervention versus knowing authorization;
47. effective-target and decision-basis reconstruction;
48. action, run, episode, and incident separation;
49. cross-run dependence;
50. recorded-reasoning limits;
51. outcome-stage separation; and
52. derivative-source treatment.

The pilot separately reports:

53. inter-practitioner agreement;
54. alignment with the author-developed reference model; and
55. independent expert assessment of whether that model is defensible.

Scenario provenance, the construction oracle, controlled evidence degradation, recognizability review, and access history are sealed before testing. After completion, practitioners report prior familiarity and any recognized source case. Agreement is analyzed separately where exposure or recognition could have influenced the result.

Stage 1 does not establish real-world forensic reliability, external validity, legal acceptance, or general utility. Later stages require artifact-level synthetic testing, retrospective case evaluation, and external replication.

Appropriate publication claim

Before testing, HAAD may be described as:

> *A publicly released forensic and assurance methodology for determining how far available evidence supports attribution of a consequential AI-agent action to a natural person. Version 0.5 remains subject to structured testing and external review.*

After a successful Stage 1 pilot, the bounded claim is:

> *Independent practitioners reached materially consistent results when applying HAAD Version 0.5 to the tested structured synthetic scenarios, and independent review found the ceiling-setting reference conclusions defensible.*

That is not equivalent to saying HAAD has been validated as a standard or established as reliable for all forensic contexts.