Human–Agent Attribution Discontinuity

Executive Overview of HAAD Version 0.4

Version 0.4 · Public-review and structured-testing draft

Author: Kevin V. Watson

DOI: 10.5281/zenodo.21683276

Parent method: The Zemi Method

Canonical source: Human–Agent Attribution Discontinuity: A Forensic and Assurance Methodology, Version 0.4

Notice. Public-review and structured-testing draft. Not a validated standard, forensic doctrine, or legal attribution framework.

Contents

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:

- a precise proposition;
- supporting and contrary evidence;
- an articulated inference and warrant;
- materially plausible competing origins;
- a proposition-specific convergence analysis;
- an adjudication status; and
- 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.

What HAAD produces

HAAD produces:

1. an attribution graph showing the material causal and evidentiary relationships;
2. proposition-level adjudications;
3. a discontinuity schedule identifying affected dimensions and failure conditions;
4. a bounded investigative-attribution conclusion;
5. a statement of the precise point at which human attribution stops; and
6. 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.

Separate attribution dimensions

HAAD separates:

- **AD-01 External action**;
- **AD-02 Technical actor and execution**;
- **AD-03 Credential or account**;
- **AD-04 Human identity**;
- **AD-05 Natural-person control**;
- **AD-06 Authorship of the material interaction**;
- **AD-07 Instruction origin**;
- **AD-08 Knowing authorization**;
- **AD-09 Voluntariness**;

- **AD-10 Scope**;
- **AD-11 Knowledge**;
- **AD-12 Intent**;
- **AD-13 Delegation**; and
- **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.

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:

1. a **plausibility gate**—technical feasibility plus a case-specific evidentiary or contextual anchor; and
2. 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.

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.

Two-axis discontinuity classification

HAAD separates:

- **Affected attribution dimension:** what material proposition failed.
- **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.

Proportional application

HAAD supports three application levels:

- **Triage:** provisional action, attribution path, evident discontinuities, competing origins, and boundary.
- **Standard:** full graph, evidence and warrant analysis, competing-origin testing, adjudication, discontinuity schedule, and bounded conclusion.
- **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 0.4 is prepared for structured-scenario testing.

As of July 29, 2026, neither Round 1 nor Round 2 has been conducted with independent practitioners. Version 0.4 revisions are based on review and scenario analysis, not reported empirical pilot findings.

The Stage 1 pilot evaluates:

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

The pilot separately reports:

1. inter-practitioner agreement;
2. alignment with the author-developed reference model; and
3. independent expert assessment of whether that model is defensible.

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 public-review and structured-testing draft of a forensic and assurance methodology for determining how far available evidence supports attribution of a consequential AI-agent action to a natural person.

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

> Independent practitioners reached materially consistent results when applying HAAD Version 0.4 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.