Human–Agent Attribution Discontinuity

A Forensic and Assurance Methodology

Version 0.4 · Public-review and structured-testing draft

Date: July 29, 2026

Author: Kevin V. Watson

DOI: 10.5281/zenodo.21683276

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

Parent method: The Zemi Method (Watson, 2026a)

Inherited companion: The Shared Evidentiary Spine for AI Systems, Version 0.2

Version 0.4 companions: HAAD Executive Overview; HAAD Practitioner Workpapers; HAAD Structured-Scenario Pilot Evidence Packs; and the sealed HAAD Structured-Scenario Pilot Reference Model and Agreement Instrument

Material changes from Version 0.3

Reframed testing claims as formative structured-scenario evaluation; established a canonical fourteen-dimension register; added Authorship of the material interaction and Interaction-authorship discontinuity; reconciled the taxonomy, workflow, conclusion layers, workpapers, pilot, and reference model to that register; strengthened Delegated Risk Tolerance safeguards; renamed and made proposition-specific the independent-convergence rule; operationalized material plausibility; distinguished authorization from scope; defined primary and contributing discontinuities; separated attribution dimensions from failure conditions; added applicability dispositions and testimonial-proposition discipline; added proportional application levels and a parallel organizational attribution path; revised workpapers and quality controls; incorporated the lineage of inter-agent instruction laundering; and established a two-round external testing design.

Publication status notice. This publication-development version is released for practitioner review and structured scenario testing. It has not been established as a validated standard, forensic doctrine, legal attribution framework, or generally reliable method for real-world casework. Pilot agreement demonstrates specification clarity and inter-practitioner consistency only within the tested materials.

Contents

1. Purpose

This methodology determines whether the available evidence supports the attribution of an AI agent's consequential action to a natural person, their authority, knowledge, or intent. It identifies and classifies every material break or weakness in that attribution.

It addresses a recurring inferential error:

> The agent acted through an account associated with a person; therefore, that person knowingly caused or authorized the action.

The premise may establish a technical association. It does not, without further evidence, establish human control, knowing authorization, intent, or responsibility.

The methodology can be used in two modes:

- **Retrospective forensic mode:** beginning with an observed action or consequence and determining how far attribution can defensibly proceed toward a natural person.
- **Prospective assurance mode:** beginning with a proposed consequential agent capability and determining whether the system would produce evidence sufficient to support or resist later human attribution.

The method produces a bounded attribution account, a discontinuity finding where a material transition is not established, and a record of what additional evidence would be required to resolve it.

Its claimed contribution is a forensic and assurance methodology for determining how far available evidence supports attribution of a consequential AI-agent action to a natural person. The contribution lies in proposition-level attribution analysis, separation of attribution dimensions, graph-based location of evidentiary breaks, proposition-specific source-dependence analysis, bounded conclusions, and conversion of retrospective gaps into readiness requirements. HAAD does not claim to be a complete theory of AI responsibility.

2. Relationship to the existing methodology architecture

This is a focused attribution diagnostic, not a replacement for AI Incident Reconstruction, AI Governance Claim Validation, or AI Forensic Readiness Assessment.

- **AI Incident Reconstruction** establishes what happened. This methodology tests whether and how the reconstructed action can be attributed across the human–agent boundary.
- **AI Governance Claim Validation** tests claims such as "every consequential action is attributable to an authorized human." This methodology supplies the attribution-specific propositions and evidence expectations.
- **AI Forensic Readiness Assessment** asks whether a future incident could be reconstructed and defended. This methodology supplies the human-attribution evidence requirements for that assessment.

It inherits *The Shared Evidentiary Spine for AI Systems*, Version 0.2: observation is separated from inference; the subject system's records are treated as an untrusted witness until corroborated; data

may operate as instruction across sessions; silence is evidence; the examiner's tools are validated; and conclusions are bounded as known, assumed, or undetermined (Watson, 2026b).

2.1 Relationship to Artifact-to-Finding Promotion

Artifact-to-Finding Promotion is the author's related concept for the controlled movement of an AI-generated artifact—a lead, output, association, or risk signal—into a human-owned finding. The promotion is legitimate only where a human adjudicator can state the supporting observations, inference, warrant, uncertainty, and accountability for the conclusion.

The two concepts address opposite evidentiary directions:

```
AI artifact → finding about the world
Agent-linked artifact → finding about a person
```

HAAD identifies where the second movement lacks continuity. A misattribution occurs when an agent-linked artifact, account, credential, signature, or trace is promoted into a finding about a natural person beyond what the evidence supports. This definition makes the relationship internal and explicit; it does not ask the reader to rely on an unidentified external document.

2.2 Established foundations and claimed contribution

HAAD does not claim to originate structured hypothesis comparison, digital-evidence preservation, provenance modeling, prompt-injection analysis, or the philosophical responsibility gap.

The requirement to test materially plausible alternative origins is adapted from the logic of Analysis of Competing Hypotheses: evidence is assessed across competing explanations rather than collected only in support of a preferred one (Heuer, 1999). HAAD narrows that discipline to attributional transitions in human–agent systems.

The preservation, integrity, chain-of-custody, contemporaneous-documentation, and source-validation requirements follow established digital-forensic guidance (Kent et al., 2006; SWGDE, 2025). Provider-dependent acquisition and the possibility that cloud evidence is unavailable, incomplete, proprietary, or timestamp-distorted are recognized in cloud-forensics guidance (SWGDE, 2024).

The graph's use of entities, activities, agents, derivations, and temporal relationships is compatible with the W3C PROV family, although HAAD adds a forensic adjudication layer that PROV does not itself supply (Moreau & Missier, 2013).

The data-as-instruction problem draws on prompt-injection work identifying the structural danger of processing trusted instructions and untrusted content without a reliable boundary (Willison, 2022; OWASP Foundation, n.d.). HAAD treats that security condition as an attribution problem when external content, retrieved data, or stored state may be the effective origin of an agent action.

The mechanism by which relayed content acquires authority its origin does not support is documented in the multi-agent security literature. Indirect prompt injection blurs the boundary between data and instruction (Greshake et al., 2023). Malicious content can propagate across agents through ordinary communication channels (Lee & Tiwari, 2024). Intermediate agents can reformat adversarial instructions into trusted-looking outputs, an effect prior work has itself called laundering (Triedman et al., 2025; Jha et al., 2026), and agents may treat peer agents as trusted sources (Lupinacci et al., 2025). HAAD does not claim this phenomenon or the laundering

metaphor. Its contribution is the evidentiary treatment: distinguishing originating and immediate sources at each relay, recording trust representations as propositions, and bounding natural-person attribution when knowing and voluntary human conduct is not established.

The responsibility-gap literature asks whether responsibility can fairly attach when autonomous-system behavior exceeds human prediction or control (Matthias, 2004). HAAD asks the prior forensic question: what evidence establishes identity, control, authorization, voluntariness, knowledge, intent, and causal contribution before a responsibility inquiry begins?

Emerging Internet-Drafts propose human-anchored agent identity and cryptographic delegation provenance (Beyer, 2026; Helixar, 2026). They are cited as works in progress, not adopted standards. HAAD's distinct claim is that verifiable technical provenance still requires proposition-level adjudication before it supports attribution to a natural person's knowing and voluntary conduct.

3. Defined concept

3.1 Human–Agent Attribution Discontinuity

Human–Agent Attribution Discontinuity (HAAD) is a break, weakness, or unresolved inferential transition in one or more of the canonical attribution dimensions connecting an AI agent's action to a natural person. The canonical dimensions are defined in Section 5.3.

A discontinuity does not require a total absence of records. It can exist where records are abundant but establish only technical events, accounts, credentials, declared ownership, or nominal delegation.

3.2 Attributional transition

The unit of analysis is the **attributional transition**: a proposition that one evidentiary object or actor is linked to the next for a specified purpose.

Examples:

- The external transaction was caused by the recorded tool call.
- The tool call was issued by the identified agent instance.
- The agent instance operated in the identified session.
- The credential was exercised by the identified device or session.
- The natural person controlled that device or session at the material time.
- The person authored the specific material interaction attributed to them.
- The effective instruction originated in that interaction rather than another source.
- The person knowingly authorized the action represented by the interaction.
- The later action remained within the authority that person granted.
- The consequential method or outcome was within the person's knowledge or intent.

Each transition is adjudicated separately. Continuity is never inferred merely because the endpoints are associated.

3.3 Material transition

A transition is **material** when the final attribution conclusion depends on it. A missing cosmetic field is not necessarily a discontinuity. A missing link between a credential and the person alleged to have controlled it is.

3.4 Technical continuity and human continuity

Technical continuity exists when evidence connects systems, agent instances, sessions, credentials, tool calls, and external actions.

Human continuity exists only to the extent evidence connects a natural person to control, authorship of the material interaction, knowing participation, authorization, or intent.

Technical continuity does not create human continuity by implication.

3.5 Delegated Risk Tolerance

Delegated Risk Tolerance (DRT) is a documented statement of the categories, severity, limits, and conditions of risk that a person knowingly accepts when delegating discretionary action to an agent.

DRT is an evidentiary object, not a waiver and not a conclusion. It may help establish awareness and accepted scope where the record shows:

- the record was sufficiently specific and intelligible to the person;
- the person had actual authority to accept the stated delegation;
- the agent's material capabilities were accurately disclosed;
- the foreseeable risk classes were disclosed;
- the included and excluded targets, values, methods, durations, and consequences were stated;
- the accepted degree of agent discretion was stated;
- the conditions for confirmation, escalation, and termination were stated;
- the disclosures and record remained current at the material time;
- no intervening model, configuration, interface, policy, or capability change invalidated the disclosure;
- acceptance was affirmative and distinguishable from passive acceptance of standard terms;
- material limits were implemented as enforceable controls, or the absence of enforcement was disclosed and evaluated; and
- the eventual action remained within the documented boundaries.

A generic acceptance of autonomous operation does not establish knowledge of every possible action, authorization of conduct outside the stated tolerance, or intent concerning a particular consequence.

A DRT records the boundaries of disclosed and accepted delegation. It does not transfer responsibility for system design, capability changes, control enforcement, logging, monitoring, or organizational risk acceptance to the individual. A signed DRT may be stronger evidence of

organizational notice of a risk or control requirement than of individual authorization or responsibility for a later action.

3.6 Inter-agent instruction laundering

Inter-agent instruction laundering is a process in which untrusted, unresolved, or lower-trust content is interpreted, summarized, transformed, or relayed by one agent and then received by another agent as if it carried greater authority or trust than its origin supports.

The term describes an evidentiary transformation. It does not require the agents to possess consciousness, moral agency, or an intention to deceive.

---

4. Scope and boundaries

4.1 What the methodology does

It:

- decomposes a human-attribution claim into testable propositions;
- maps the relevant action, identity, delegation, instruction, and control relationships;
- tests the integrity, completeness, provenance, and independence of the supporting evidence;
- evaluates competing explanations for every material transition;
- identifies the exact point at which attribution ceases to be established;
- distinguishes attribution of identity, control, authorization, knowledge, intent, causation, and responsibility;
- supports a bounded investigative attribution or a documented non-attribution; and
- converts evidentiary gaps into assurance and forensic-readiness requirements.

4.2 What the methodology does not do

It does not:

- equate authenticated activity with natural-person authorship;
- equate technical permission with substantive authority;
- infer intent merely from consequences;
- assign moral blame, civil liability, criminal liability, or legal responsibility;
- certify that cryptographic provenance is complete or truthful;
- reconstruct an incident where the underlying action sequence has not first been established; or
- guarantee future attribution merely because controls passed a point-in-time assessment.

The highest conclusion reached is investigative attribution. Legal attribution remains for the appropriate adjudicative body.

4.3 Proportional application

HAAD is applied in proportion to consequence, contestability, uncertainty, and intended use:

- **Triage application:** records the consequential action, proposed attribution, material causal path, evident discontinuities, competing origins that are immediately material, and a provisional attribution boundary.
- **Standard application:** completes the attribution graph, proposition-specific evidence and warrant analysis, competing-origin testing, transition adjudication, discontinuity schedule, and bounded conclusion.
- **Enhanced application:** completes all workpapers and adds formal preservation, examiner-tool validation, expanded alternative-origin testing, organizational-path analysis, and independent technical or forensic review. It is expected for high-impact, contested, regulatory, employment, disciplinary, insurance, civil, or criminal matters.

Every level retains the narrow-proof rule, transition decomposition, materially plausible competing-origin analysis, and a bounded conclusion. Triage does not support a final adverse finding about a person and must be labelled provisional. Escalate to Standard or Enhanced application when a proposed conclusion may materially affect a person's rights, employment, reputation, access, legal position, or financial interests.

5. Governing principles

5.1 No transitive attribution

Attribution does not pass automatically through a chain.

If an agent is linked to a credential and the credential is registered to a person, the agent's action is not thereby attributed to that person. The credential-to-person transition requires its own evidence.

5.2 The narrow-proof rule

Every artifact is credited only with what it proves.

- A valid signature proves the use of a key under the applicable trust model.
- An authentication event proves that a mechanism accepted presented factors.
- An account record proves the system's association between an account and an identity.
- A delegation token proves the recorded grant and its integrity, if validated.
- An agent trace proves what the preserved trace records, subject to integrity and completeness.

None, alone, proves natural-person control, voluntariness, knowledge, intent, or the completeness of the surrounding record.

5.3 The attribution-dimension separation rule

The following register is canonical throughout HAAD. These dimensions are distinct propositions and must not be collapsed:

ID	Canonical attribution dimension	Question	Section 7 classification
AD-01	External action	What consequential event occurred?	Action discontinuity (7.1)

ID	Canonical attribution dimension	Question	Section 7 classification
AD-02	Technical actor and execution	Which tool, process, agent instance, and material execution sequence caused it?	Execution discontinuity (7.2)
AD-03	Credential or account	Which account, token, key, service identity, or credential enabled the material operation?	Identity discontinuity (7.3)
AD-04	Human identity	Which natural person, if any, is supported as associated with the material credential, device, session, or interaction?	Identity discontinuity (7.3)
AD-05	Natural-person control	Did that person control the relevant credential, device, or session at the material time?	Control discontinuity (7.4)
AD-06	Authorship of the material interaction	Did that person produce, adopt, or knowingly cause the specific instruction, approval, click, message, or decision attributed to them?	Interaction-authorship discontinuity (7.5)
AD-07	Instruction origin	What source introduced the effective instruction or material premise on which the agent acted?	Instruction-origin discontinuity (7.6)
AD-08	Knowing authorization	What action did the person knowingly permit?	Authorization discontinuity (7.7)
AD-09	Voluntariness	Was the person's participation or authorization voluntary rather than compelled, coerced, or produced under duress?	Voluntariness discontinuity (7.8)
AD-10	Scope	Did the agent's conduct remain within the permitted objective, means, targets, limits, and duration?	Scope discontinuity (7.9)
AD-11	Knowledge	What material facts, capabilities, risks, and constraints were presented to or known by the person?	Knowledge discontinuity (7.10)
AD-12	Intent	What action or consequence did the person purposefully seek or knowingly adopt?	Intent discontinuity (7.11)
AD-13	Delegation	What authority passed through human, agent, or operator relationships, and on what conditions?	Delegation discontinuity (7.12)
AD-14	Causal contribution	Which instruction, state, intervention, actor, or decision materially contributed to the action?	Causal-contribution discontinuity (7.17)

The first fourteen rows of the Step 10 and WP-08 conclusion registers reproduce this order. **Individual investigative attribution** and **organizational investigative attribution** are synthesized conclusions drawn from the dimension findings; they are not additional dimensions. **Responsibility referral** is a downstream referral output identifying which established and unresolved findings must be passed to a competent organizational, regulatory, or legal decision-maker; HAAD does not decide what responsibility attaches.

Establishing one dimension does not establish another. Where two dimensions map to the same Section 7 classification, as Credential or account and Human identity do under Identity discontinuity, the finding must still name the precise canonical dimension affected.

5.4 The nearest-established-actor rule

Attribution proceeds only as far as the evidence permits. When a material discontinuity is reached, the practitioner stops and attributes the action only to the nearest actor or system state that is established.

For example:

> The payment is established as an action of Agent A operating under service credential C. The evidence does not establish which natural person controlled the initiating session. Human attribution is undetermined.

Stopping is a valid methodological result.

5.5 The weakest-material-transition ceiling

The overall attribution conclusion cannot be stronger than its weakest material transition. Strong evidence elsewhere cannot compensate for a missing transition essential to the conclusion.

Evidence concerning a downstream attribution dimension cannot cure an unestablished upstream transition necessary to connect that evidence to the proposed person. Evidence that an action was intentional, for example, does not establish whose intent it was where identity, control, or authorship of the material interaction remains discontinuous.

5.6 The competing-origin rule

An apparent human instruction is tested against materially plausible alternative origins, including:

- authorized direct human instruction;
- use of another person's authenticated session;
- stolen or delegated credentials;
- an authenticated and genuine controller acting under coercion, compulsion, or duress;
- malware or remote control;
- current-session prompt injection;
- a delayed trigger from persistent memory or retrieved content;
- agent-generated reinterpretation or plan expansion;
- agent-to-agent deception or synthetic-context manipulation;
- inter-agent instruction laundering, where an untrusted origin gains apparent authority through relay or transformation;

- sub-agent or operator intervention;
- policy, system-prompt, or orchestration influence;
- autonomous action within broad technical permissions; and
- incomplete, altered, or fabricated provenance.

An explanation is not included merely because it is imaginable. It passes the **plausibility gate** when it is technically feasible in the material architecture and is anchored by at least one case-specific indicator, known compromise or control failure, observed inconsistency, credible source, applicable threat condition, or unresolved pathway that the available evidence was reasonably expected to exclude. It passes the **materiality gate** when accepting it could change a material transition, affected attribution dimension, or attribution boundary. Both gates are documented. Architectural possibility or potential effect on the conclusion, standing alone, is insufficient.

5.7 The proposition-specific independent-convergence rule

Known attribution of natural-person control or knowing authorization ordinarily requires convergence among at least two proposition-relevant evidentiary streams with materially distinct capture or failure paths. Two records produced by the same identity or logging pipeline may constitute one stream for one proposition while independently proving a different proposition.

Independence is a property of capture and failure paths, not vendor count. Sources operated by one provider may supply meaningful convergence where they observe the event through distinct mechanisms and do not share the material manipulation path. Sources operated by different providers may remain dependent where one merely reproduces an assertion received from the other.

Independence dimension	Test
Causal independence	Did each source observe the material event, or did one copy, inherit, or transform the other's assertion?
Control independence	Could the same actor, agent, credential, administrator, or control plane alter both?
Technical independence	Do the sources use distinct capture, transmission, storage, integrity, and time mechanisms?
Administrative independence	Are the sources governed by distinct roles, custodians, retention controls, or trust domains?

Where two fully independent streams are not reasonably available, the practitioner documents the acquisition efforts, source coverage, dependencies, common failure risks, and materially plausible competing explanations; seeks the strongest available convergence; and states the resulting conclusion ceiling. The result does not automatically become undetermined merely because one provider operates the ecosystem. Stream count is not a substitute for quality.

A single-stream proposition may be known—established only through a documented exception showing:

1. why a second proposition-relevant stream is unavailable, unnecessary, or not reasonably obtainable;
2. whether the stream directly observed the proposition or recorded, copied, or transformed another source's assertion;

3. the stream's integrity, completeness, reliability, and material-time coverage;
4. every actor, credential, administrator, agent, or control plane capable of altering it;
5. what confirming or contradictory records the system was expected to produce and whether they exist;
6. the materially plausible alternatives and the stream's capacity to resist them;
7. the residual common failure or manipulation risk; and
8. why that residual risk does not require a lower classification or narrower conclusion.

In Enhanced application, a second practitioner must expressly concur with a single-stream known—established classification. The exception is proposition-specific and does not elevate the same stream for another attribution dimension.

5.8 The time-bounded attribution rule

Identity, control, permissions, memory, model configuration, and delegation are evaluated at the time of the material action. Ownership or configuration before or after the event does not establish the state at the event.

5.9 Organizational attribution is a separate path

Failure to attribute an action to an individual does not automatically establish organizational attribution. Organizational attribution requires its own propositions concerning deployment, granted authority, control ownership, notice, and governance decisions.

The individual and organizational paths may be evaluated in parallel, but neither is a fallback assumption.

5.10 No presumed singular origin

The method does not require every action to have one root human, orchestrator, or instruction. Where peer agents, several humans, external sources, or policy components jointly contribute, the practitioner maps causal contribution without forcing the graph into a hierarchy. A finding may establish plural contribution while leaving any single originating cause undetermined.

5.11 Delegated discretion does not collapse attribution dimensions

A person's knowing acceptance that an agent may select tools or methods autonomously can establish awareness of agent discretion. It does not, without more, establish:

- knowledge of a capability or risk class that was not disclosed;
- authorization outside the accepted risk, value, target, method, or duration boundary;
- knowledge of a later agent adaptation;
- voluntariness;
- intent concerning a particular action or consequence; or
- responsibility for everything technically possible within the system.

Where informed acceptance of the relevant risk class is established but the eventual action exceeded the accepted boundary, the discontinuity ordinarily lies in scope rather than knowledge. Where the capability or risk class was not adequately disclosed, knowledge remains discontinuous.

5.12 Primary and contributing discontinuities

The primary discontinuity is the earliest material transition, along the causal path relevant to the scoped attribution proposition, whose failure independently prevents the attribution dimension in issue. “Earliest” refers to causal order on that path, not merely timestamp or visual graph position.

Other classifications are reported as contributing, causal, or downstream discontinuities. A failure-condition qualifier does not displace the affected attribution dimension as the primary finding. When several transitions independently set the same ceiling, report co-primary discontinuities and explain why no single transition controls.

5.13 Authorization and scope selection rule

Use **Authorization discontinuity** where no valid grant covers the category of consequential action. Use **Scope discontinuity** where a valid grant covers the category, but the action exceeds a target, recipient, amount, method, time, duration, objective, tool, onward delegation, or other limiting condition.

Both may be reported. Authorization is ordinarily primary when the action category itself was never granted. Scope is ordinarily primary when the category was granted but a limiting condition was exceeded; action-specific authorization is then reported as unestablished to the extent that it depends on the exceeded boundary.

5.14 Testimonial-proposition discipline

An interview, declaration, or recorded statement establishes that the statement was made when its authenticity is established. It does not automatically establish the truth of every proposition asserted.

The practitioner decomposes admissions, denials, failures of recollection, explanations, and exculpatory assertions; identifies the speaker's interest in the attribution outcome; and evaluates contemporaneity, specificity, internal consistency, external consistency, and independent corroboration. An unsupported denial ordinarily leaves the underlying proposition undetermined. It becomes known—contradicted only when reliable evidence affirmatively establishes that the proposed transition did not occur as alleged.

5.15 Organizational attribution path

Organizational investigative attribution is evaluated through its own propositions and graph:

```
Organizational decision-maker or governing body
        ↓
Deployment and capability decision
        ↓
Authority, permission, and control design
        ↓
Organizational risk acceptance and notice
        ↓
Monitoring, enforcement, and change management
        ↓
Agent capability and operation
        ↓
Consequential action
```

The path tests who selected and deployed the system, configured authority, accepted known risks, owned material controls, approved capability changes, received notice, and monitored or enforced

limits. It supports organizational investigative attribution and readiness findings; it does not decide corporate liability or serve as an automatic fallback when individual attribution fails.

6. Attribution graph

The practitioner constructs an **attribution graph**, not merely a linear chain. Agentic actions may involve multiple humans, agents, sub-agents, peer agents, credentials, stored memories, retrieved sources, tools, policy gates, and external systems.

A minimum graph contains:

```
Natural person or unresolved human origin
        ↓
Device, session, or interaction
        ↓
Credential and authenticated identity
        ↓
Material interaction and its authorship
        ↓
Instruction, authorization, or adopted objective
        ↓
Agent instance and configuration
        ↓
Delegation, memory, retrieval, and sub-agent branches
        ↓
Tool invocation
        ↓
External action
        ↓
Consequence
```

The arrows are propositions, not decoration. Each receives an identifier, evidence, competing explanations, and a Zemi classification.

The graph also records:

- **forks**, where one instruction produces several agent courses of action;
- **merges**, where several human, system, or environmental inputs contribute to one action;
- **transformations**, where an agent materially interprets, expands, or changes an instruction;
- **trust transformations**, where relayed content acquires, loses, or obscures an authority or trust label;
- **interventions**, where another actor or control changes the course;
- **stored-state crossings**, where an input from an earlier session becomes causally active later; and
- **external boundaries**, where evidence passes to a different platform, operator, or trust domain.

The graph must declare its observed topology:

- **hierarchical**, where authority or work descends from an identifiable principal or orchestrator;

- **peer-to-peer**, where agents exchange tasks or state without a controlling root visible in the occurrence;
- **mesh**, where several agents repeatedly influence one another;
- **hybrid**, where hierarchical delegation and peer interaction coexist; or
- **undetermined**, where the surviving evidence cannot establish the topology.

For peer-to-peer and mesh systems, the practitioner records contribution paths, feedback loops, and intervention points. The method does not manufacture an originating human or agent merely to complete a tree.

Where one agent supplies context to another, the graph distinguishes the originating source of the material content, the relaying or transforming agent, the representation made to the receiving agent, the trust or authority status attached at each hop, the immediate source that influenced the action, and the originating source that introduced the material premise or instruction.

---

7. Discontinuity taxonomy

A finding is classified on two axes:

- **Axis 1—affected attribution dimension:** which canonical AD-01 through AD-14 proposition failed.
- **Axis 2—failure condition:** why continuity failed or cannot be demonstrated. Defective provenance, temporal ambiguity, external intervention, evidence access, dependent corroboration, compromised source, contradiction, or another stated evidentiary condition.

Sections 7.1 through 7.12 and 7.17 classify affected dimensions. Sections 7.13 through 7.16 identify common failure conditions. The canonical register in Section 5.3 controls the dimension name; Section 7 supplies the corresponding discontinuity classification. A matter may contain several dimensions and conditions, but the report does not force them into competing labels. State the finding in combined form where useful, for example:

> Authorization discontinuity caused by dependent provenance and unresolved intervention.

The primary-discontinuity rule in Section 5.12 controls the headline finding. Failure conditions and downstream affected dimensions remain visible as contributing classifications.

7.1 Action discontinuity

The external consequence is observed, but its causal connection to the alleged agent action is not established.

7.2 Execution discontinuity

The agent's involvement is established, but the material sequence of model, orchestration, memory, sub-agent, and tool events cannot be reconstructed.

7.3 Identity discontinuity

The technical actor is not reliably linked to the material credential or account, or the material credential, account, device, or declared owner is not reliably linked to a natural person. The finding names whether AD-03 Credential or account, AD-04 Human identity, or both are affected.

7.4 Control discontinuity

A natural person is associated with the credential, device, or account, but the evidence does not establish that the person controlled it at the material time.

Control continuity establishes control of the relevant credential, device, or session. It does not by itself establish authorship of every instruction, approval, click, message, or decision recorded during the interval.

7.5 Interaction-authorship discontinuity

A natural person is established as controlling a relevant credential, device, or session, but the evidence does not establish that the person produced, adopted, or knowingly caused the specific instruction, approval, click, message, or other material interaction attributed to them.

This classification marks the transition from general control to the particular human act asserted. It does not require proof that no human interaction occurred; it records that authorship by the proposed person is not established.

7.6 Instruction-origin discontinuity

The agent acted on an effective instruction, but its source cannot be established. This includes unresolved data-as-instruction and delayed-trigger cases.

7.7 Authorization discontinuity

The agent possessed technical permission, but the evidence does not establish substantive human authorization for the action.

7.8 Voluntariness discontinuity

The natural person is authenticated, present, and in control, but the evidence does not establish that their participation or authorization was voluntary. Coercion, compulsion, duress, or comparable constraint remains materially plausible.

This is not an identity or control failure. It limits what may be inferred from genuine human participation.

7.9 Scope discontinuity

An authorization existed, but the evidence does not establish that the agent's selected objective, means, target, value, duration, or downstream delegation remained inside it.

7.10 Knowledge discontinuity

Human participation is established, but the evidence does not establish what material capabilities, risks, constraints, or later adaptations the person knew.

Knowledge of general autonomy is not knowledge of every available capability. The practitioner evaluates DRT, interface disclosures, warnings, capability descriptions, prior experience, and contemporaneous communications to determine whether the relevant capability and risk class were presented and understood.

If the person knowingly accepted autonomous tool selection within a disclosed risk class, knowledge of agent discretion may be established. Whether the particular action remained within

the accepted boundary is then adjudicated under Scope Discontinuity. Particular intent remains separate.

7.11 Intent discontinuity

The person is linked to the interaction, but the evidence does not establish that the consequential action or outcome fell within their purpose or knowing adoption.

7.12 Delegation discontinuity

Authority passed through one or more agents or operators, but a material delegation, attenuation, transformation, or recipient cannot be established.

7.13 Provenance discontinuity

Records exist but their lineage, completeness, integrity, trust boundary, or relationship to the occurrence cannot be established.

7.14 Temporal discontinuity

The relevant events cannot be reliably ordered, or material state changed across sessions without preserved snapshots sufficient to connect cause and action.

7.15 Intervention discontinuity

The evidence cannot resolve whether another human, agent, injected source, compromised component, or policy mechanism redirected the operation.

This includes agent-to-agent deception, synthetic-context manipulation, and inter-agent instruction laundering. The practitioner does not attribute human voluntariness or moral agency to the machines. The forensic question is whether one agent introduced or relayed materially false, misleading, untrusted, or instruction-bearing content that redirected another agent, and whether the originating and immediate sources can be distinguished.

7.16 Evidence-access discontinuity

Evidence material to a transition may exist within a third-party provider, operator, proprietary system, or foreign jurisdiction, but is not available to the examiner. The finding distinguishes:

- **absent evidence**, which was not created, was not retained, or no longer exists;
- **inaccessible evidence**, which may exist but cannot be obtained within the examiner's authority, technical access, provider capability, legal process, time, or scope; and
- **evidence of unknown existence**, where reasonable inquiries cannot establish whether the material evidence exists.

Provider silence is not treated as proof that the evidence does not exist. The acquisition steps attempted, preservation requests, provider representations, legal or contractual restrictions, and unresolved production questions are recorded.

7.17 Causal-contribution discontinuity

Several actors, agents, or sources materially contributed to an action, but the evidence cannot resolve their relative or necessary causal contribution. The method reports the contribution graph and does not force plural causation into a single-origin attribution.

In an inter-agent manipulation case, the immediate cause may be the receiving agent's reliance on synthesized context, while the originating cause lies in another agent, a retrieved source, poisoned memory, or a human instruction. Both contribution positions are recorded rather than collapsed.

8. Evidence taxonomy

Evidence is collected by the proposition it can support, not merely by system component.

8.1 Action and outcome evidence

External service records, transaction ledgers, recipient records, network observations, file-system changes, repository history, communications, physical outcomes, and independently generated receipts.

8.2 Agent and execution evidence

Agent-instance identifiers, model and version, system prompt, configuration, orchestration traces, tool calls, sub-agent messages, intermediate artifacts, policy decisions, retries, failures, and execution environment state.

8.3 Identity and control evidence

Account records, authentication events, identity-provider records, token issuance and exchange, key-use records, endpoint telemetry, device possession, session binding, IP and network context, remote-access records, malware evidence, and contemporaneous human activity.

Records from a common provider are not presumed either independent or dependent. Their capture mechanisms, inherited fields, control planes, custodians, and common failure paths are documented.

Evidence of general device or session control is distinguished from evidence of authorship of a material interaction. Interaction-authorship evidence may include independently captured input events, action-specific approval receipts, user-bound signatures over the material content, screen or interface-state capture, contemporaneous communications, and other records connecting the person to the precise instruction, approval, click, message, or decision.

8.4 Instruction and intent evidence

Prompts, structured requests, approvals, signed intent or delegation records, user-interface displays, confirmation events, warnings shown, conversation history, capability disclosures, risk disclosures, DRT records, changes to objectives, cancellations, corrections, and contemporaneous communications.

The recorded prompt is evidence of recorded content. It is not automatically evidence of authorship, complete intent, voluntariness, or awareness of later agent adaptations.

8.5 Authority and scope evidence

Policies, role assignments, capability grants, delegation chains, approval requirements, contractual or organizational authority, limits on value, target, method, time, tool, and onward delegation, plus revocation and exception records.

8.6 Causal-context evidence

Retrieved documents, browser content, emails, files, environmental observations, persistent memory, session summaries, vector-store entries, system messages, guardrail interventions, operator changes, and sources capable of functioning as instruction.

8.7 Integrity and completeness evidence

Hashes, signatures, trusted timestamps, append-only or write-once properties, schema definitions, coverage tests, dropped-event counters, clock synchronization, retention settings, access-control history, export procedures, and evidence of who could alter each source.

---

9. Practitioner workflow

Step 1. Frame the attribution question

State the consequential action, the person or organization to whom attribution is proposed, the attribution dimensions actually in issue, the period, systems, and exclusions.

Select and justify the Triage, Standard, or Enhanced application level. If organizational attribution is in scope, identify its decision, authority, control, notice, monitoring, and enforcement propositions separately from the individual path.

Do not begin with "Who is responsible?" Begin with propositions capable of evidentiary testing.

Example:

> Does the evidence establish that Person P controlled Session S and knowingly authorized Agent A to execute Transaction T on Recipient R for Amount V?

Step 2. Establish the action before attributing it

Establish that the external action or consequence occurred and distinguish it from the agent's own claim that it occurred. Identify the first independently reliable record of the action.

If the action itself is not established, report an action discontinuity and do not proceed as though the consequence were proven.

Step 3. Construct the attribution graph

Map every material actor, system, credential, instruction, delegation, stored-state source, tool, action, and consequence. Assign an identifier to every material transition.

Mark trust-domain crossings and identify which sources share a causal or logging origin. In multi-agent matters, record content transformations and trust-status changes at every material relay.

Step 4. Decompose the proposed attribution

Convert the overall attribution into propositions, one per material transition and attribution dimension. State the expected evidence for each proposition before evaluating the collected evidence.

Example propositions:

1. Transaction T was caused by Tool Call TC-1.

2. TC-1 was issued by Agent Instance A-17.
3. A-17 operated within Session S-9.
4. Credential C-4 authorized S-9.
5. Person P controlled S-9 at the material time.
6. Person P authored Material Interaction MI-2.
7. Effective Instruction I-2 originated in MI-2 rather than in another human, retrieved source, stored state, agent transformation, or policy component.
8. Person P knowingly authorized T's amount, recipient, method, and timing.
9. No material intervening source redirected A-17.

Step 5. Preserve and test the evidence

Preserve volatile session, memory, model, identity, token, and orchestration state. Record chain of custody.

For each source, test:

- authenticity;
- integrity;
- completeness for the proposition;
- timestamp reliability;
- retention and temporal coverage;
- ability of the subject to alter the record;
- causal independence from other sources;
- inherited or copied fields, common control planes, and shared failure paths; and
- whether the record captures failed, suppressed, modified, and successful events.

A source that fails does not disappear. Its limitation becomes part of the finding. If a provider-controlled source cannot be obtained, record whether the evidence is known absent, reasonably believed to exist but inaccessible, or of unknown existence. Follow provider-dependent acquisition and verification practices appropriate to the authority and matter (SWGDE, 2024).

For an all-in-one platform, complete a source-dependence analysis rather than treating the provider as either one indivisible source or several independent sources by default. Determine what each record observed directly, which fields it inherited, who could alter it, which control plane governed it, and what compromise or failure could make the records wrong in the same way.

Step 6. Record observations separately from inferences

Record what each artifact shows on its face. Then state the inference proposed from it and the warrant connecting the two.

Example:

- **Observation:** The identity provider recorded successful authentication for Account P at 09:14 using a registered factor.
- **Inference proposed:** Person P controlled the session.

- **Required warrant:** Evidence that the factor and session were under P's control and were not delegated, stolen, remotely operated, or replayed.

Step 7. Test competing explanations

For each material transition, identify the materially plausible alternatives and specify what evidence supports, contradicts, or cannot resolve each one.

The purpose is not to invent doubt. It is to prevent the preferred attribution from receiving a lower evidentiary burden than its alternatives.

Apply both gates in Section 5.6. Record the case-specific anchor that makes each alternative plausible and the transition or boundary it could change. Exclude alternatives supported only by abstract technical possibility or imagined doubt.

Where an agent received material context from another agent, test whether the content was a direct observation, inference, instruction, fabrication, or transformation; whether its origin and trust status survived the relay; and whether the receiving agent treated it as more authoritative than the evidence warranted.

Step 8. Adjudicate each transition

Classify each proposition using the Zemi discipline:

- **Known—established:** reliable evidence establishes the transition and materially plausible alternatives have been addressed.
- **Known—contradicted:** reliable evidence establishes that the proposed transition did not occur as alleged.
- **Assumed:** the transition is used for a limited analytical purpose but is not established; the assumption and its effect are explicit.
- **Undetermined—pending:** potentially resolvable evidence remains to be obtained or tested.
- **Undetermined—exhausted, absent:** the necessary evidence was not created, was not retained, was destroyed, or is established not to exist.
- **Undetermined—exhausted, inaccessible:** the evidence may exist, but reasonable acquisition avenues within the examiner's authority, scope, and available process have been exhausted.
- **Undetermined—exhausted, existence unknown:** reasonable inquiries cannot establish whether the necessary evidence exists.

The following are scope dispositions, not evidentiary conclusions:

- **Not in issue:** the dimension belongs to the conceptual model, but no material fact, context, or competing explanation places it in dispute.
- **Not applicable:** the dimension does not logically apply to the scoped attribution claim.

Neither disposition means known—established. Do not require proof of the absence of every imaginable condition. For example, voluntariness becomes an adjudicated proposition when evidence or context makes coercion, compulsion, duress, or comparable constraint materially plausible.

Move an undetermined proposition from pending to an exhausted state only after recording the evidence sought, systems and custodians checked, time and scope coverage, acquisition steps and authority, provider or witness responses, and why further reasonable inquiry is unavailable or disproportionate. A missing expected record may support non-demonstration or a control failure; it does not by itself establish that no other evidence exists.

Do not assign one confidence label to the whole graph before adjudicating its propositions.

Step 9. Identify and classify each discontinuity

A discontinuity is reported where a material transition is contradicted, assumed, or undetermined.

For each discontinuity, state:

- canonical AD identifier and affected dimension name;
- failure condition or conditions;
- whether the discontinuity is primary, co-primary, contributing, causal, or downstream;
- location in the graph;
- evidence available;
- reason continuity is not established;
- competing explanations that remain;
- effect on the overall conclusion;
- whether resolution is pending or exhausted; and
- evidence or control that would have prevented or resolved it.

Select the primary discontinuity under Section 5.12 and distinguish Authorization from Scope under Section 5.13.

Step 10. Set the attribution boundary

Apply the nearest-established-actor rule and weakest-material-transition ceiling. State exactly how far the attribution proceeds and where it stops.

Separate conclusions for:

- external action;
- technical actor and execution;
- credential or account;
- human identity;
- natural-person control;
- authorship of the material interaction;
- instruction origin;
- knowing authorization;
- voluntariness;
- scope;
- knowledge;

- intent;
- delegation;
- causal contribution;
- individual investigative attribution;
- organizational investigative attribution;
- responsibility referral.

Step 11. Report and return assurance findings

Issue the bounded attribution conclusion, discontinuity schedule, unresolved alternatives, limitations, and readiness implications.

Every retrospective discontinuity is also considered prospectively:

> What instrumentation, identity binding, authorization design, retention, or independent record would allow this transition to be tested in a future matter?

---

10. Discontinuity determination

The methodology does not use a single numerical score in Version 0.4. A score could hide the location and consequence of a broken transition and imply that strength elsewhere offsets a material gap.

The determination is proposition-based.

10.1 No demonstrated discontinuity

Every material transition necessary for the stated attribution dimension is known—established on evidence of sufficient integrity and independence.

This does not mean no discontinuity exists outside the defined scope.

10.2 Bounded discontinuity

A material transition is assumed or undetermined, but the discontinuity affects only a specified attribution dimension or portion of the graph.

Example: technical attribution to an agent is established; human intent is undetermined.

10.3 Attribution-breaking discontinuity

A contradicted, assumed, or undetermined transition prevents the proposed human-attribution conclusion.

Example: the credential is linked to Person P, but control of the material session is undetermined. The action cannot be attributed to P as its human controller.

10.4 Misattribution finding

Reliable evidence contradicts a human attribution that was nevertheless asserted, operationalized, or used as the basis for a consequential decision.

This finding connects HAAD directly to Artifact-to-Finding Promotion: an agent-linked artifact was improperly promoted into a finding about a person.

11. Sufficiency rules

A material transition may be known—established only when:

1. the proposition is precise;
2. the evidence relied upon is authentic and sufficiently intact;
3. the evidence covers the material time and scope;
4. observation and inference are separated;
5. the inference has an articulated warrant;
6. materially plausible competing explanations have been tested;
7. the conclusion does not exceed what each source narrowly proves; and
8. for natural-person control or knowing authorization, proposition-relevant independent evidentiary streams ordinarily converge;
9. testimonial assertions are not treated as proof of their contents without the analysis required by Section 5.14; and
10. any single-stream exception satisfies the documented requirements in Section 5.7.

The transition remains assumed or undetermined when any necessary condition is absent.

Absence of evidence is not automatically evidence that a transition did not occur. It is evidence of non-demonstration unless the system was reliably expected to produce the artifact and that expectation itself is established.

12. Standard workpapers

The separately packaged *HAAD Practitioner Workpapers, Version 0.4* reproduces these forms and adds operational control sheets. This methodology remains canonical if the extracted pack differs.

WP-01 Attribution Question and Scope Record

Field	Entry
Consequential action	
Proposed human or organizational attribution	
Attribution dimensions in issue	
Period and material time	
Systems and trust domains	
Included questions	
Excluded questions	
Mode: forensic or assurance	
Application level and justification	

Field	Entry
Dimensions not in issue	
Dimensions not applicable	
Individual path, organizational path, or both	

WP-02 Attribution Graph and Transition Register

Transition ID	From	To	Proposition	Material to which conclusion	Expected evidence	Trust-boundary crossing	Topology or contribution role
T-01							

WP-03 Evidence Integrity, Completeness, and Independence Record

Complete one row for each proposition–stream relationship. The same record may be independent for one proposition and dependent or irrelevant for another.

Proposition	Stream ID	Evidence and independent face fact	Integrity and completeness	Capture mechanism	Dependencies or inherited fields	Control plane or custodian	Common failure path	Evidentiary contribution and limit
	S-01							

Where a single-stream known—established exception is proposed, append:

Proposition	Reason second stream is unavailable or unnecessary	Direct observation or inherited assertion	Alteration capability	Expected confirming or contradictory records	Alternatives tested	Residual failure risk	Why classification is not lowered	Second-practitioner concurrence, if required

WP-04 Observation, Inference, and Warrant Log

Entry	Observation	Proposed inference	Warrant required	Support for warrant	Assumption or limit
OI-01					

WP-05 Competing-Origin and Intervention Matrix

Transition ID	Hypothesis	Plausibility anchor	Material effect if accepted	Originating source	Immediate influencing source	Content or trust transformation	Supporting evidence	Contrary or missing evidence	Status
T-01									

WP-06 Transition Adjudication Matrix

Transition ID	Proposition	Evidence relied upon	Proposition-specific convergence	Zemi classification or scope disposition	Exhaustion basis, if applicable	Reason	Effect on attribution
T-01							

WP-07 Discontinuity Schedule

HAAD ID	Canonical dimension: AD ID and name	Failure condition	Role: primary, co-primary, contributing, causal, or downstream	Graph location	Resolution status	Remaining alternatives	Conclusion ceiling	Required future evidence
H-01								

WP-08 Bounded Attribution Conclusion

Conclusion layer	Finding	Classification or scope disposition	Attribution boundary and basis
External action			
Technical actor and execution			
Credential or account			
Human identity			
Natural-person control			
Authorship of the material interaction			
Instruction origin			
Knowing authorization			
Voluntariness			
Scope			
Knowledge			
Intent			
Delegation			
Causal contribution			
Individual investigative attribution			
Organizational investigative attribution			
Responsibility referral			

WP-09 Assurance and Readiness Return

Discontinuity	Evidence that was absent or unreliable	Control or instrumentation requirement	Owner	Priority	Reassessment trigger
H-01					

WP-10 Organizational Attribution Path

Complete for Enhanced application and whenever organizational attribution is in scope.

Transition ID	Organizational actor or function	Decision, authority, notice, control, or enforcement proposition	Evidence	Classification	Effect on organizational attribution	Readiness implication
OT-01						

13. Reporting language

Use language that identifies both the established endpoint and the stopping point.

13.1 Technical continuity established; human continuity undetermined

> The evidence establishes that Agent A, operating in Session S under Credential C, executed Action X. The available evidence does not establish which natural person controlled Session S at the material time. Attribution to the registered owner of Credential C would exceed the evidence.

13.2 Human control established; authorization scope undetermined

> The evidence establishes that Person P controlled the initiating session and authorized Objective O. It does not establish that Method M or Consequence X fell within the scope P granted. Human control is established; authorization of the consequential action is undetermined.

13.3 Intent discontinuity

> Person P is established as the author of the initiating instruction. The agent independently selected the material target and method after subsequent retrieval and planning. The evidence does not establish that P knew of or intended those selections.

13.4 Voluntariness discontinuity

> The evidence establishes that Person P controlled the session and issued Authorization A. Evidence material to whether that authorization was voluntary is unavailable, and coercion remains a materially plausible explanation. Identity, control, and the recorded authorization are established; voluntary authorization is undetermined.

13.5 Delegated discretion established; scope exceeded

> The evidence establishes that Person P knowingly authorized autonomous tool selection within Risk Tolerance R. The relevant capability and risk class were disclosed. Agent A selected Tool T, but its target and transaction value exceeded the limits recorded in R. Knowledge of agent discretion is established; authorization of the consequential action is not. The attribution discontinuity lies in scope.

13.6 Platform monoculture

> The identity, application, orchestration, and API records are operated by Provider V. Source-dependence analysis establishes that the identity record and external API receipt observed distinct events through separate capture and control paths, while the application and orchestration records share inherited identifiers and a common logging plane. The four records are therefore treated as two evidentiary streams, not one merely because they share a provider and not four merely because they are separate logs.

13.7 Inter-agent instruction laundering

> Agent B acted on a representation relayed by Agent A. The evidence establishes that the material instruction originated in untrusted retrieved content, was summarized by Agent A without its trust limitation, and was presented to Agent B as orchestration context. Agent A was the immediate source; the retrieved content was the originating source. Human authorization of the relayed instruction is not established.

13.8 Misattribution

> The organization attributed Action X to Person P on the basis that P was the registered owner of Credential C. The evidence establishes credential association but does not establish P's control of the material session. The human-attribution finding was therefore promoted beyond the status supported by the evidence.

13.9 Non-attribution

> The investigation establishes the action and technical actor but reaches an attribution-breaking discontinuity at the session-control transition. Individual human attribution is undetermined—exhausted on the available evidence. This is a bounded non-attribution, not a finding that no human participated.

13.10 Device use established; authorship of the material interaction undetermined

> The evidence establishes that Person P controlled and actively used Device D during the material interval. The subject platform records Interaction I during that interval, but no independent evidence connects P's device activity to that specific interaction. Device control is established; authorship of Interaction I and any authorization inferred from it remain undetermined.

14. Quality controls

A second practitioner checks that:

- the action was established before human attribution began;
- the overall claim was decomposed into transition-level propositions;
- the graph includes relevant branches, merges, interventions, stored state, and trust-domain crossings;
- technical identity was not treated as natural-person identity;
- account ownership was not treated as session control;
- general credential, device, or session control was not treated as authorship of the material interaction;
- technical permission was not treated as substantive authorization;
- authorization was not treated as knowledge or intent;
- Authorization and Scope were selected under the category-versus-limiting-condition rule;
- genuine control or recorded authorization was not treated as proof of voluntariness;
- every artifact was credited only with its narrow proof;
- materially plausible alternatives passed both the plausibility and materiality gates;

- two purportedly independent sources do not share the same causal origin;
- evidentiary-stream independence and count were assessed for a named proposition;
- every single-stream known—established exception contains the required showing and, in Enhanced application, express second-practitioner concurrence;
- source independence was adjudicated through capture and failure paths rather than vendor count;
- a platform monoculture was not automatically treated as either one source or several independent sources;
- DRT was tested for specificity, intelligibility, authority, disclosure accuracy, currency, affirmative acceptance, tolerance boundaries, and operational enforcement;
- a DRT was not treated as a waiver or transfer of organizational control responsibility;
- acceptance of autonomous tooling was not treated as acceptance of every technically possible action;
- interview statements were decomposed and interested or exculpatory assertions were not credited as self-proving;
- inter-agent relays preserved or disclosed origin and trust status, or the resulting discontinuity was reported;
- originating and immediate influencing sources were not collapsed in instruction-laundering cases;
- affected attribution dimensions were distinguished from failure-condition qualifiers;
- the primary discontinuity was selected on the scoped causal path;
- not-in-issue and not-applicable dispositions were not treated as established findings;
- any exhausted status was supported by a documented inquiry and source-coverage record;
- no conclusion exceeds its weakest material transition;
- individual and organizational attribution were adjudicated separately;
- assumptions and undetermined states were not hidden inside a global confidence label; and
- the report states where attribution stops.

A second practitioner using the same preserved evidence and rules should identify the same material transitions, discontinuities, and attribution boundary. Disagreement is documented at the proposition level.

15. Limitations

- Human voluntariness, knowledge, and intent are rarely recorded directly and commonly require careful inference from conduct and context.
- Strong authentication reduces some alternative explanations but does not eliminate coercion, delegation, compromise, remote control, or incomplete intent.
- Privacy, employment, privilege, contractual, and cross-border restrictions may prevent collection of evidence relevant to human control.
- Agent state may be volatile, summarized, overwritten, or distributed across providers.

- Multi-agent systems may produce causal contribution rather than a single originating cause.
- A documented non-attribution may be the strongest correct result.
- The methodology evaluates evidentiary sufficiency; it does not determine the legal standard applicable in a particular jurisdiction.

16. Structured-scenario testing program

16.1 Purpose and limits

Stage 1 is a formative evaluation of specification clarity, usability, and inter-practitioner consistency on structured synthetic scenarios. It does not test acquisition, preservation, authentication, chain of custody, timestamp analysis, examiner-tool reliability, source completeness, performance on raw artifacts, external validity, forensic reliability, legal acceptability, or real-world utility.

Agreement may show that practitioners understood and applied the tested specification consistently. It does not establish that their shared conclusion is substantively correct outside the scenario or that HAAD is a validated method.

16.2 Independent evaluation roles

Report three comparisons separately:

1. **Inter-practitioner agreement:** whether practitioners independently identify materially equivalent transitions, classifications, evidentiary streams, and attribution boundaries.
2. **Reference-model alignment:** whether each practitioner's result aligns with the sealed author-developed reference model.
3. **Independent defensibility review:** whether an external reviewer who did not develop the method, scenarios, or reference model considers the reference model and any materially different practitioner answer defensible.

At least one independent reviewer must have had no role in developing the method, scenario, evidence pack, or reference model. The reviewer should produce an independent analysis before seeing the sealed reference model.

16.3 Two-round design

Empirical status at release: As of July 29, 2026, Round 1 has not been conducted with independent practitioners. Version 0.4 corrections are review- and scenario-analysis-driven. They are not reported as empirical pilot findings.

Round 1—under-specification discovery. Practitioners apply the frozen Version 0.3 materials. Predicted divergence-log items and unpredicted disagreements are recorded at proposition level. The purpose is formative: identify ambiguity, usability problems, and uncontrolled interpretation.

Revision. Version 0.4 incorporates explicit rules only where the evidence from review, scenario analysis, or Round 1 supports correction. The change log distinguishes defect correction from preference.

Round 2—operational reproducibility. New practitioners who did not participate in Round 1 apply Version 0.4 to unseen or materially altered scenarios. Reusing the same facts as the author-developed reference cases is insufficient. Round 2 is the stronger test of whether the revised method produces stable results.

Round 2 scenario titles and instructions must not reveal the expected discontinuity. The set should include at least one delegated-risk case with only one genuinely ambiguous limiting condition, one multi-agent case in which lineage must be reconstructed rather than stated in the inventory, one shared-provider case that separates device use from authorship of the material interaction, and one control case in which every material transition within scope is sufficiently established. At least one scenario should be authored or independently challenged by a person who did not develop HAAD.

16.4 Stage 1 acceptance criterion

The Version 0.4 specification passes the Stage 1 structured-scenario pilot only if new practitioners applying unseen or materially altered scenarios:

1. identify the same ceiling-setting transitions;
2. identify the same affected attribution dimensions;
3. distinguish primary discontinuities from contributing failure conditions;
4. count independent streams consistently for each named proposition;
5. stop human attribution at the same evidentiary boundary; and
6. distinguish retrospective attribution from prospective readiness findings; and
7. correctly reach no demonstrated discontinuity when the evidence establishes every material transition within scope.

Secondary-label or incidental-grouping differences are recorded but do not constitute boundary failure. Differences between pending and an exhausted access state at the same stopping point are recorded as exhaustion-state divergences, not boundary disagreements. Results are reported by measure and proposition and are not averaged into one score.

In addition, the independent reviewer must rate every ceiling-setting reference proposition and attribution boundary as defensible after completing an analysis without access to the sealed author model. A not-defensible rating prevents a Stage 1 pass even if practitioners agree with one another.

Passing this criterion supports only a claim of inter-practitioner consistency within the tested scenario class. It does not establish external validity or general forensic reliability.

16.5 Later evaluation stages

- **Stage 2—artifact-level synthetic testing:** practitioners acquire, preserve, authenticate, and analyze synthetic source artifacts rather than stipulated inventory entries.
- **Stage 3—retrospective case evaluation:** suitably governed testing on closed or reconstructed matters, with comparison against independently established case facts where available.
- **Stage 4—field and external replication:** independent teams apply the method in different systems and domains, report usability and disagreement, and test whether results remain bounded and reproducible.

17. Version 0.4 research questions

Scenario testing should answer:

1. Can practitioners identify material transitions consistently without the graph becoming unmanageably detailed?
2. Does proposition-specific independent-convergence analysis operate consistently where one provider controls several capture systems?
3. Are knowledge and intent inside the same methodology operationally useful, or should they become a separate interpretive module?
4. Can organizational attribution be evaluated in parallel without being treated as an automatic fallback?
5. Does two-axis classification distinguish affected attribution dimensions from failure conditions consistently in practice?
6. What minimum evidence supports natural-person control in passwordless, shared-device, delegated-assistant, and remote-administration environments?
7. Does the contribution-path treatment distinguish originating, transforming, immediate, and acting sources in peer-to-peer and mesh systems without implying a false hierarchy?
8. When does an unresolved discontinuity become exhausted rather than pending?
9. Can practitioners distinguish informed acceptance of a risk class from authorization or intent concerning a particular consequential action?
10. Do practitioners distinguish device control from authorship of a specific material interaction?
11. Do application levels reduce administrative burden without weakening the minimum attribution safeguards?
12. Does two-axis classification reduce disputes between affected dimensions and failure conditions?
13. Can an independent reviewer distinguish author-model alignment from substantive defensibility?
14. Can practitioners apply the single-stream exception consistently without using it as a routine substitute for convergence?
15. Can practitioners correctly reach no demonstrated discontinuity without relaxing the transition-level sufficiency rules?

References

Beyer, B. W. (2026). *Architecture for human-anchored agent identity, delegation, and provenance* (Internet-Draft draft-beyer-agent-identity-architecture-00). Internet Engineering Task Force. Work in progress. https://www.ietf.org/archive/id/draft-beyer-agent-identity-architecture-00.html

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In

Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79–90). ACM. https://doi.org/10.1145/3605764.3623985

Helixar. (2026). *Human Delegation Provenance Protocol (HDP): Cryptographic chain-of-custody for agentic AI systems* (Internet-Draft draft-helixar-hdp-agentic-delegation-00). Internet Engineering Task Force. Work in progress. https://www.ietf.org/archive/id/draft-helixar-hdp-agentic-delegation-00.html

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency. https://www.cia.gov/resources/csi/books-monographs/psychology-of-intelligence-analysis-2/

Jha, R., Triedman, H., Wagle, J., & Shmatikov, V. (2026). Breaking and fixing defenses against control-flow hijacking in multi-agent systems. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*. https://openreview.net/forum?id=PNU9Rj5RDQ

Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to integrating forensic techniques into incident response* (NIST Special Publication 800-86). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-86

Lee, D., & Tiwari, M. (2024). *Prompt infection: LLM-to-LLM prompt injection within multi-agent systems* (arXiv:2410.07283). arXiv. https://doi.org/10.48550/arXiv.2410.07283

Lupinacci, M., Pironti, F. A., Blefari, F., Romeo, F., Arena, L., & Furfaro, A. (2025). *The dark side of LLMs: Agent-based attacks for complete computer takeover* (arXiv:2507.06850). arXiv. https://arxiv.org/abs/2507.06850

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Moreau, L., & Missier, P. (Eds.). (2013). *PROV-DM: The PROV data model* (W3C Recommendation). World Wide Web Consortium. https://www.w3.org/TR/prov-dm/

OWASP Foundation. (n.d.). *LLM prompt injection prevention cheat sheet*. OWASP Cheat Sheet Series. Retrieved July 28, 2026, from https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html

Scientific Working Group on Digital Evidence. (2024). *Best practices for digital evidence acquisition, preservation, and analysis from cloud service providers* (SWGDE 23-F-004-1.1, Version 1.1). https://www.swgde.org/documents/published-complete-listing/23-f-004-best-practices-for-digital-evidence-acquisition-preservation-and-analysis-from-cloud-service-providers/

Scientific Working Group on Digital Evidence. (2025). *Best practices for digital evidence collection* (SWGDE 18-F-002-2.0, Version 2.0). https://www.swgde.org/documents/published-complete-listing/18-f-002-best-practices-for-digital-evidence-collection/

Triedman, H., Jha, R. D., & Shmatikov, V. (2025). Multi-agent systems execute arbitrary malicious code. In *Proceedings of the Second Conference on Language Modeling (COLM 2025)*. https://openreview.net/forum?id=DAozI4etUp

Watson, K. V. (2026a). *The Zemi Method* (Version 1.0). https://zemimethod.com

Watson, K. V. (2026b). *The Shared Evidentiary Spine for AI Systems* (Version 0.2, working draft). Unpublished manuscript.

Willison, S. (2022, September 17). You can't solve AI security problems with more AI. *Simon Willison's Weblog*. https://simonwillison.net/2022/Sep/17/prompt-injection-more-ai/

Public-review and structured-testing draft. The name and transition-level diagnostic are proposed here; the evidentiary and analytical foundations are attributed. Version 0.4 strengthens classification, convergence, delegated-risk, proportionality, organizational-path, and testing controls. It is not a validated standard or doctrine. External review and staged testing should determine which distinctions hold.