Human–Agent Attribution Discontinuity

A Forensic and Assurance Methodology

Version 0.5

Kevin V. Watson

August 25, 2026

Document information

Version 0.5 DOI: 10.5281/zenodo.22099968

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

Status: Public release. 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 0.5 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*

The pilot evidence packs and sealed reference model are testing-design instruments, not completed practitioner results or empirical outputs.

Material changes from Version 0.4

Added a protocol-mediated attribution application profile for Model Context Protocol (MCP) systems; defined the host-to-downstream transition chain; added protocol, authorization, consent, task-state, trace-correlation, tool-definition, implementation, and downstream-execution evidence requirements; added MCP-specific reporting language and readiness controls; added rules for later constraints and operative state, subject-system representations, capability-authority-enforcement separation, compound incidents, multiple downstream actions, layered actor and sponsorship paths, confidence transfer, activity-share claims, human decision gates, campaign aggregation, positive authorization, approval and silence, recorded reasoning, cross-run state persistence, outcome-stage separation, intra-organizational decomposition, technical-actor identity layers, effective-target reconstruction, decision-basis state, and derivative-source treatment; added pilot controls for scenario provenance, recognition contamination, construction oracles, controlled evidence degradation, recognizability review, participant exposure, and sensitivity reporting; added WP-11 and WP-12 plus structured scenarios addressing these conditions; clarified that protocol identity, authenticated client access, approval events, cryptographic receipts, structured incident entries, and technique mappings do not by themselves establish natural-person control, interaction authorship, knowing authorization, voluntariness, intent, responsibility, or independent corroboration; and preserved the canonical fourteen-dimension register unchanged.

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 and actor paths, graph-based location of evidentiary breaks, proposition-specific source-dependence analysis, operative-state reconstruction, 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, control-flow hijacking can propagate unsafe action across an orchestration path, and agents may treat peer-agent output as trusted context (Triedman et al., 2025; Jha et al., 2026; Lupinacci et al., 2025). HAAD uses **inter-agent instruction laundering** as a proposed evidentiary label for the resulting trust transformation. The label describes the change in apparent trust and does not imply deliberate concealment or an intention to deceive. Its contribution is the treatment of originating and immediate sources at each relay, trust representations as propositions, and bounded 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.

The Model Context Protocol standardizes communication among host applications, MCP clients, MCP servers, and exposed tools. Core authorization, tool-discovery, and elicitation exchanges can produce records material to attribution. The separately versioned MCP Tasks extension can add durable task-state records where both parties adopt it. Trace correlation, downstream receipts, and other observability records remain implementation or external-system evidence rather than proof supplied automatically by the core protocol. Protocol continuity is therefore not human continuity. HAAD treats MCP as a technology-specific application profile rather than a new canonical attribution dimension and tests what each record establishes, where state or authority changed, and whether the record can be correlated to the consequential downstream action (Model Context Protocol, 2026a, 2026b, 2026c, 2026d, 2026e, 2026f).

The Decision Evidence Maturity Model identifies a container fallacy in treating the presence of logs, traces, or other evidence containers as if it established their sufficiency for a governance question (Solozobov, 2026). HAAD accepts that broader problem and addresses a narrower object: whether proposition-level evidence supports attribution across technical, individual, organizational, and sponsorship paths. DEMM classifies reconstructability and evidence maturity by property. HAAD classifies attributional transitions, discontinuities, and stopping boundaries. The methods are complementary, and HAAD does not claim to originate the general proposition that record presence is not evidentiary sufficiency.

The AI Incident Database is a structured incident and source index, not a substitute for the records or reports it aggregates (AI Incident Database, 2026). MITRE ATT&CK campaign and technique mappings can organize reported activity, but a mapping derived from one provider's report does not independently corroborate that provider's factual or actor-attribution conclusions (MITRE, 2026). HAAD

treats registry entries, taxonomy labels, and technique mappings as source-attributed analytic artifacts whose dependencies must be preserved.

The Agents of Chaos exploratory red-teaming study supplies realistic examples involving persistent memory, multi-party communication, tool use, non-owner instructions, destructive actions, and discrepancies between system statements and external state (Shapira et al., 2026). It is a scenario and artifact seed rather than a population estimate or a sealed reference model. The aiAuthZ proposal places identity-bound authorization in a separate trust domain and produces authenticated messages, policy decisions, and receipts (Kodathala, 2026). Those controls may strengthen technical attribution and enforcement, but their artifacts remain subject to HAAD's narrow-proof rule. Key use, message binding, policy acceptance, and receipt generation do not alone establish natural-person control, authorship, voluntariness, understanding of effective targets, or knowing authorization of every downstream consequence.

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 imply deliberate concealment and 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

Section 5 states the controlling attribution principles. The practitioner workpapers and testing companions operationalize these principles but do not replace, narrow, or expand them. If an operational instruction conflicts with this specification, the specification controls and the conflict is recorded as a correction candidate.

FOUNDATIONAL CONTINUITY AND EVIDENCE RULES

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.
- A consent record proves the recorded scope grant and its integrity, if validated. It does not, by itself, prove knowing authorization, understanding of the effective parameters, or voluntariness.
- An approval event proves that the recorded interface or control accepted the captured response, if validated. It does not, by itself, prove who controlled the response mechanism, what the person understood, whether the display matched the effective action, or whether non-intervention was knowing authorization.
- A cryptographic authorization receipt proves the recorded binding among the key, message, policy decision, and receipt under the applicable trust model, if validated. It does not, by itself, prove natural-person key control, message authorship, voluntariness, understanding, or authorization of an undisclosed transformation or consequence.
- An agent trace proves what the preserved trace records, subject to integrity and completeness.
- A recorded reasoning trace proves that the text or structured reasoning record appeared in the preserved trajectory, subject to integrity, completeness, and architecture. It does not self-prove a faithful causal explanation, subjective awareness, belief, intent, deception, or the completeness of the system's internal computation.
- A subject-system statement proves that the statement was generated or recorded, subject to integrity and attribution. It does not, by itself, prove that the stated event occurred, that the stated explanation is accurate, or that a claimed test, restoration, limitation, reason, or mental state is true.
- A provider confidence label proves the provider reported that assessment. It does not transfer the provider's confidence to the practitioner without access to the material evidence, method, source weighting, and competing-explanation analysis supporting it.
- An incident-registry entry proves that the registry recorded and classified reports under its procedures. It does not independently prove the reported events or make the linked reports independent evidentiary streams.
- A taxonomy or technique mapping proves that the analyst or catalog assigned the recorded labels. It does not independently establish the underlying activity, actor, campaign boundary, sponsorship, or completeness of the mapped sequence.

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)
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)

ID	Canonical attribution dimension	Question	Section 7 classification
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.

The ceiling is proposition- and path-specific. A weak organizational or state-sponsorship transition limits that conclusion without automatically reducing a separately established technical-execution finding. Conversely, strong technical attribution does not elevate an unestablished human, organizational, or state-attribution path.

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.

The operative state also includes later instructions, freezes, revocations, narrowed approvals, component and package versions, effective code, tool definitions, policy state, and downstream configuration. A later constraint changes the recorded attribution boundary only to the extent that its authorship or authority, delivery, timing, applicability, and continued operation at the material transition are established. Continued technical capability does not establish continued authority.

LAYERED ATTRIBUTION AND APPLICATION RULES

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.

5.16 Capability, authority, and enforcement separation

Technical capability, recorded authority, and control enforcement are separate propositions. The fact that a credential, agent, tool, or service could perform an action does not establish that the action was authorized. The absence or failure of an enforcement control does not establish knowing authorization, although it may support a separate readiness or organizational-control finding.

Retrospective attribution and prospective readiness are reported separately. A historical finding that an action exceeded a valid limit does not depend on whether the limit was technically enforced. A recommendation for stronger enforcement does not prove what occurred historically.

5.17 Subject-system representation rule

A representation produced by the system under examination is treated as a subject-system statement. It may establish that the representation occurred and may contribute to reconstruction when independently supported. It does not self-prove the represented event, the success of a test, the availability of restoration, the completeness of records, the cause of an action, or any claimed mental state, reason, intent, concealment, or inability.

Materially inaccurate, incomplete, contradictory, or self-exculpatory representations are adjudicated as separate propositions from the underlying action. Reports should describe the representation and its inconsistency without attributing deception, panic, concealment, or comparable purpose unless evidence supports the relevant human intent.

5.18 Compound-incident and multiple-action decomposition

Where an incident includes several actions, attempted remediations, substitute data, secondary disclosures, tests, reports, or representations, each materially consequential event is framed and adjudicated separately. Common timing, platform, agent identity, tool invocation, or workflow does not establish common purpose, causal continuity, authorization, or intent.

One authorized invocation may produce several downstream actions. Authorization of the stated action does not establish authorization of an undisclosed addition, transformed parameter, secondary recipient, exfiltration, destructive side effect, or later generated representation.

5.19 Component identity and implementation-integrity rule

Vendor identity, product or package name, repository similarity, publisher account, protocol identity, and natural-person authorship are separate propositions. None alone establishes vendor authorization, package provenance, publisher control, implementation integrity, or the person who authored or deployed material code.

A tool description or schema establishes the function represented to the model or client. It does not establish that the invoked handler, package version, effective configuration, or downstream request conformed to that representation. Where material, compare the arguments authorized or transmitted at every transition with the effective request received by the downstream service, and record every material addition, omission, transformation, or secondary action.

5.20 Source-attributed assessment and confidence-transfer rule

An external analytic conclusion is decomposed into the propositions it asserts and the evidence access available to the practitioner. Where the underlying evidence or method is unavailable, report the conclusion as an assessment of the named source rather than adopting its confidence label as the practitioner's own.

An evidence-access discontinuity does not establish that the source assessment is wrong. It limits independent reproduction and requires the report to distinguish:

- facts directly established from accessible records;
- source-attributed assessments supported by inaccessible or undisclosed material;
- independently corroborated conclusions; and
- propositions that remain assumed or undetermined.

5.21 Activity-share and strategic-decision-gate rule

A percentage of operations, requests, elapsed time, tool calls, tokens, or estimated effort attributed to an AI system is not itself an attribution dimension and does not allocate knowledge, intent, authorization, causation, or responsibility. Any quantitative autonomy claim must state its denominator, unit of analysis, measurement method, source coverage, exclusions, and uncertainty.

Human intervention at a target-selection, escalation, exploitation, credential-use, lateral-movement, release, publication, exfiltration, retention, or comparable decision gate may be materially significant even when automated systems perform most tactical operations. The practitioner maps each gate and tests who authored, controlled, and knowingly authorized the transition.

5.22 Episode and campaign aggregation rule

Findings are bounded to the actions, sessions, targets, components, periods, and evidentiary coverage actually examined. A result established for one tool call, host, victim, successful action, or sampled episode is not generalized to a campaign, organization, or population without a stated aggregation rule and evidence of completeness or representative selection.

Campaign-level conclusions must identify the unit of analysis, inclusion and exclusion criteria, successful and unsuccessful episodes, coverage gaps, common and divergent execution paths, and whether the evidence supports aggregation. Where those conditions are not met, report proposition-level or episode-level findings and identify the campaign-wide claim as source-attributed, assumed, or undetermined.

PRESSURE-TESTED SPECIALIZED APPLICATION RULES

5.23 Positive authorization, approval, and silence rule

The absence of an express prohibition does not establish authorization. It also does not, by itself, establish that an action was prohibited. The practitioner identifies the positive grant, the authority from

which it arose, its objectively established limits, and whether those limits were operative at the material transition.

An approval click, command confirmation, continued session, failure to interrupt, or other non-intervention is adjudicated separately from knowing authorization. The record must establish what action, target, parameters, state basis, risk, and consequence were presented; who controlled the response mechanism; and whether the effective action conformed to that presentation. Permission to execute a command string does not establish knowing authorization of a materially different effective target or consequence.

5.24 Recorded-reasoning rule

Recorded model reasoning, summaries, scratchpads, plans, explanations, and self-reports are subject-system representations. They may support temporal, instruction-lineage, and competing-explanation analysis when correlated with preserved context, tool calls, and external actions. They do not establish a human-like mental state or a faithful account of the computation that caused the action.

Where a report uses terms such as awareness, belief, deception, concealment, or intent, the practitioner states whether the term describes observable conduct, recorded text, a source-attributed assessment, or an independently supported inference. Machine wording is not promoted into human or organizational intent.

5.25 Cross-run state persistence and independence rule

Repeated runs, sessions, or agent instances are not treated as independent where they share persistent external state, accounts, artifacts, credentials, instructions, infrastructure, operators, public messages, memory, or another causal influence capable of producing the same result. The practitioner records the shared state, direction of influence, affected runs, and limit this places on replication, frequency, and population-level conclusions.

Counts of actions, runs, episodes, and incidents are reported separately. Several actions may belong to one run, several runs may form one connected episode, and one incident may contain several episodes. Common platform or model identity does not resolve those units.

5.26 Attempt, execution, completion, effect, harm, and recovery rule

Attempt, external execution, successful completion, resulting technical or human effect, demonstrated harm, mitigation, and recovery are separate propositions. A prevented consequential action may include completed intermediate actions. The absence of an ultimate compromise or demonstrated harm does not negate those actions, and their occurrence does not establish the unrealized outcome.

Recovery does not retroactively negate deletion, interruption, disclosure, or other effects. A restoration record establishes only the restored state and coverage it can support; it does not prove that every affected item was recovered or that no interim consequence occurred.

5.27 Intra-organizational decomposition rule

An organization is not treated as one undifferentiated mind or actor. Where material, the practitioner separately identifies governing bodies, risk owners, system selectors, developers, evaluators, operators, identity administrators, control owners, monitors, incident responders, and other functions. For each function, test authority, knowledge, decision, control, notice, intervention, and causal contribution.

Statements that an organization knew, authorized, controlled, failed, or responded must identify the function and evidence supporting the proposition. Established deployment or control ownership does not automatically establish organizational knowledge, intent, responsibility, or legal liability.

5.28 Technical-actor identity-layer rule

Provider, model family, model version, safeguard configuration, agent scaffold, orchestration framework, agent instance, session, trajectory, evaluation run, tool process, external account, and downstream executor are separate technical-identity propositions. A count associated with one model family does not establish one continuous agent, one configuration, one causal path, or independent repetitions.

The nearest-established-actor rule is applied at the most specific identity layer supported by the evidence. Higher-level product or provider labels are not substituted for missing instance, session, or execution evidence.

5.29 Effective-target and decision-basis state rule

A command, tool name, or displayed plan does not by itself establish the effective target or blast radius. Where material, reconstruct shell expansion, working directory, paths, environment variables, state files, inventories, plans, credentials, cloud account, region, workspace, configuration, package or handler state, downstream parameters, and external service interpretation.

The practitioner establishes which state file, inventory, resource model, retrieved record, or configuration informed the action; its provenance, version, freshness, completeness, and fitness for the decision; who or what selected or changed it; and whether its limitations were presented to the person. A stale or substituted decision basis may affect instruction origin, knowledge, scope, authorization, provenance, temporal continuity, and causal contribution without changing the human-authored high-level objective.

5.30 Derivative-source and structured-analytic-artifact rule

Incident registries, public postmortems, technique catalogs, taxonomies, summaries, research reconstructions, and downstream reporting are mapped to their originating evidence and analyses. Replication of a claim across derivative artifacts does not create independent convergence.

A structured analytic artifact may add useful decomposition, standard vocabulary, chronology, or technique mapping. Its added structure is evaluated separately from the factual and attribution claims it inherits. The practitioner preserves the artifact version, retrieval date, source links, editorial or machine-classification status, and any material change history.

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, mitigation records, recovery records, and independently generated receipts.

Where the subject system also reports whether the action succeeded, failed, was tested, caused harm, or could be reversed, preserve that representation separately and compare it with the external action, test, backup, restoration, and affected-party records. Record attempted, executed, completed, effective, harmful, mitigated, and recovered states separately.

8.2 Agent and execution evidence

Provider, model-family, model-version, safeguard-configuration, scaffold, orchestration, agent-instance, session, trajectory, run, tool-process, external-account, and downstream-executor identifiers; system prompts; 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. A registry, catalog, postmortem, and article that all repeat one provider report are not separate streams for the inherited propositions.

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, recorded reasoning, plans, 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. A reasoning trace is treated as recorded system output, not as a transparent account of internal computation or a human-like mental state.

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. Preserve the exact operation, target, state basis, parameters, risk, and consequence displayed at approval, the captured response, the response mechanism, the effective action, and any material difference between them.

8.6 Causal-context evidence

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

For later constraints and material decision gates, preserve the instruction or approval artifact, its author or issuing authority, delivery and acknowledgement records, the component or session that received it, its effective period, any superseding state, and evidence that it remained operative at the consequential transition.

For command-mediated or infrastructure actions, preserve the proposed command and plan together with the working directory, expansions, environment, selected state or inventory, credentials, account and region, effective API operations, and resulting resources. For repeated runs, preserve shared accounts, public artifacts, memory, infrastructure, and other persistent state capable of connecting the runs.

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.

For public or structured analytic sources, preserve the entry identifier, version or permalink, retrieval date, source links, editorial status, machine-generated classifications, change history, and dependency on originating reports. A later structured mapping may add analysis without adding an independent observation.

8.8 Protocol-mediated execution evidence

Where an action is mediated by MCP or a comparable tool protocol, preserve evidence at each material transition rather than treating the protocol exchange as one event. Relevant evidence may include:

- host application, MCP client, MCP server, model, tool, and downstream-service versions and identifiers;
- package, repository, publisher, vendor-authorization, handler-code, deployment, hash, signature, dependency, and update records needed to distinguish represented identity from effective implementation;
- the effective tool catalogue at the material time, including tool names, descriptions, input schemas, annotations, catalogue refreshes, and cached definitions;
- user and system instructions, relevant context, the model-generated tool call, parameter transformations, and the request ultimately issued by the client, server, handler, and downstream service;
- request and response bodies, protocol and correlation identifiers, timestamps, retries, errors, cancellations, resumptions, and task or state handles;
- authorization-server, token-issuance, token-exchange, subject, audience, scope, elevation, revocation, and consent records, without preserving reusable secrets in the evidentiary package;
- confirmations and elicitation presented to the person, the identity of the requesting server, the exact material parameters shown, and the response captured;
- trace context capable of correlating the host, client, server, tool, and downstream transaction, including OpenTelemetry or equivalent records where implemented;
- downstream API, transaction, repository, communications, or service records independently showing the consequential action; and

- logging coverage, clock synchronization, retention, tamper protection, and evidence of who could modify tool definitions, state, traces, or authorization records.

An authenticated MCP client may establish that a client possessed valid authority under the applicable mechanism. A tool call may establish that named arguments were transmitted. A server record may establish that a tool handler executed. None of these records, alone, 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.

Where one invocation produces several downstream actions, preserve and correlate each action separately. A consent or confirmation record covering the represented operation does not establish authorization of an undisclosed parameter, recipient, transformation, secondary disclosure, or side effect.

8.9 MCP forensic-readiness controls

Prospective assurance should determine whether the system can later support these propositions without relying on one provider's retrospective reconstruction. At minimum, assess whether the implementation:

1. assigns stable correlation identifiers across the host, MCP client, MCP server, tool, and downstream service;
2. versions or immutably records the effective tool description, schema, annotations, and catalogue state used for selection and invocation;
3. versions, authenticates, and preserves the effective package, handler code, deployment configuration, publisher identity, update path, and vendor-authorization status;
4. distinguishes natural-person input, model-generated arguments, client transformations, server transformations, handler transformations, and downstream parameters;
5. records the exact confirmation or elicitation wording, material parameters, consequence description, server identity, and response presented to the person;
6. binds task and state handles to the authorized client, user context, and material operation, and records resumptions, cancellations, and handoffs;
7. records initial scopes, step-up requests, grants, revocations, token subject and audience, and the operation that caused elevation;
8. correlates protocol records with independently administered downstream receipts or observations and detects unrepresented secondary actions;
9. synchronizes clocks, records dropped events and coverage gaps, and protects material logs against unauthorized alteration; and
10. prevents token passthrough and tests for confused-deputy, scope-inflation, state-handle hijacking, tool-definition substitution, implementation substitution, unauthorized parameter mutation, and secondary-action conditions.

Absence of a control does not establish that unauthorized conduct occurred. It defines what the evidence cannot later demonstrate and may create a prospective forensic-readiness finding.

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.

Define the unit of analysis. State whether the conclusion concerns one event, one invocation, one session, one target, a sampled episode, a campaign, an organization, or another population. For aggregated conclusions, state the inclusion, exclusion, completeness, and sampling basis before analysis.

Where repeated activity is involved, distinguish actions, runs, sessions, episodes, incidents, campaigns, and populations. State whether runs share persistent state or another causal influence and whether any claim depends on treating them as independent.

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.

Decompose materially distinct actions and representations. Attempt, external execution, successful completion, effect, harm, mitigation, recovery, substitute-data generation, test claim, restoration claim, secondary disclosure, and other downstream consequence are not treated as one event merely because they arose in the same run or invocation.

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.

Record later constraints and material state changes on the graph, including freezes, revocations, narrowed approvals, package or model updates, definition changes, configuration changes, and phase-transition approvals. Where organizational or state attribution is proposed, create separate paths for the technical executor, account operator, natural person, analytic cluster, organization, and alleged sponsor.

Separate provider, model family, model version, safeguard configuration, scaffold, agent instance, session, run, tool process, external account, and downstream executor where any distinction could change attribution. Within an organization, separate decision-makers, operators, control owners, monitors, responders, and other material functions. Connect repeated runs through any shared persistent state rather than drawing them as independent by default.

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.

For protocol-mediated actions, add propositions comparing the authorized or transmitted arguments with the parameters received by the server, handler, and downstream service. Where the conclusion relies on an external assessment, add separate propositions for the assessment's source, accessible evidence, method, confidence basis, and independent corroboration.

For approval-mediated actions, add propositions for the positive grant, approving authority, exact display, response mechanism, person controlling it, effective target, decision-basis state, material risk and consequence presented, and conformity between the approved representation and resulting action. For command-mediated actions, add propositions resolving the command against the effective environment, state, credentials, account, downstream requests, and affected resources.

Step 5. Preserve and test the evidence

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

Preserve operative instructions and constraints together with delivery, acknowledgement, applicability, and retention evidence. Preserve effective component versions, package and publisher records, handler code, deployment state, tool definitions, downstream requests, and each distinct action or representation produced. Preserve the state files, inventories, plans, working context, environment, credentials, account and region, expansions, and resource mappings needed to resolve the effective target.

Preserve reasoning traces as subject-system records and document known omissions, summarization, hidden fields, redactions, provider transformations, and whether the recorded text was available to the executing component. For repeated runs, preserve shared external artifacts, accounts, memory, infrastructure, and cross-run discoveries.

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 attempted, failed, suppressed, modified, successful, mitigated, and recovered 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.

When the source also publishes the analytic conclusion, distinguish its directly observed telemetry from its interpretation and from any identity, organizational, geographic, or sponsorship assessment. Do not count a report and the provider records summarized in that report as independent streams merely because they are separate artifacts.

Trace incident registries, catalogs, technique mappings, research reconstructions, and later reporting back to their source reports. Record what structure or analysis the derivative artifact adds and which factual or attribution propositions it merely inherits.

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.

For a subject-system or provider statement, reasoning trace, registry entry, or taxonomy mapping, record separately that the artifact exists, the proposition asserted, whether the source directly observed that proposition, what analysis it added, and what independent evidence confirms or contradicts it.

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.

Where a person approved or did not stop an action, test whether the interface and surrounding context presented the effective target, state basis, material parameters, blast radius, and consequence. Consider stale or substituted state, misleading summaries, automatic execution, fatigue, time pressure, and a materially different downstream action only where case-specific evidence passes the plausibility and materiality gates.

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.

Do not import another source's confidence label. If material supporting evidence or method is inaccessible, classify the underlying proposition under HAAD and report the external conclusion as a source-attributed assessment.

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.

Apply the ceiling separately to each material path. A technical actor, account operator, natural person, analytic cluster, organization, and alleged sponsor may each have a different stopping point and classification.

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.

State the unit of analysis and prohibit aggregation beyond the examined coverage. If an autonomy or activity-share estimate is material, report its denominator, measurement method, source coverage, exclusions, uncertainty, and the separate significance of any human decision gates.

State action, run, episode, incident, and campaign counts separately where material. Disclose shared state or dependencies that prevent repeated runs from being treated as independent. Separate attempted conduct, completed intermediate actions, ultimate effect, demonstrated harm, mitigation, and recovery.

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 0.5. 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;
11. subject-system representations and external analytic assessments are not treated as self-proving;
12. the operative instruction, constraint, component, definition, and configuration state is established at the material transition where each is necessary;
13. conclusions drawn from one event, episode, target, or sample are not generalized beyond established coverage; and
14. quantitative autonomy or activity-share claims disclose their denominator, method, source coverage, exclusions, and uncertainty;
15. an approval or non-intervention inference establishes the exact action, target, decision basis, risk, and consequence presented and compares them with the effective action;
16. the effective target is resolved through the material command, environment, state, credential, account, downstream request, and resource context;
17. decision-basis state is authenticated, versioned, time-bounded, and evaluated for freshness, completeness, and fitness where material;
18. repeated runs are treated as independent only where material shared state and common causal influences have been excluded;
19. attempt, execution, completion, effect, harm, mitigation, and recovery are not collapsed; and
20. derivative registries, taxonomies, mappings, and summaries are not counted as independent observations of inherited propositions.

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 0.5* 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	
Unit of analysis and aggregation basis	
Action, run, session, episode, incident, campaign, and population counts	
Attempted, executed, completed, effective, harmful, mitigated, and recovered states in scope	
Systems and trust domains	
Included questions	
Excluded questions	
Mode: forensic or assurance	
Application level and justification	
Dimensions not in issue	
Dimensions not applicable	

Field	Entry
Individual path, organizational path, or both	
Technical, operator, cluster, organizational, or sponsorship paths in scope	

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							

Attach an operative-state record identifying later constraints, revocations, narrowed approvals, component or definition changes, decision-basis state, effective-target context, shared cross-run state, and material human decision gates.

State ID	Instruction, constraint, component, definition, decision basis, target context, shared state, or gate	Authority or origin	Provenance, version, freshness, and fitness	Delivery or deployment evidence	Effective period	Components, runs, or actions affected	Continuity limit
ST-01							

Where approval, confirmation, silence, continued operation, or non-intervention is material, append:

Approval event	Positive grant and authority	Person or mechanism controlling response	Action, target, decision basis, parameters, risk, and consequence presented	Effective action and target	Material delta	Authorization and knowledge effect
AP-01						

WP-03 Proposition-Specific Evidence 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							

Append a derivative-source record where an incident registry, postmortem, technique catalog, taxonomy, research reconstruction, or later report repeats or transforms another source.

Derivative artifact	Version and retrieval date	Originating source or evidence	Added structure or analysis	Inherited propositions	Independent observation, if any	Stream-count treatment

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					

Append a representation record where the subject system or an external source asserts that an action occurred, a test passed, restoration was impossible, a component was legitimate, an actor was identified, or another material proposition was true.

Representation ID	Source and statement	Proposition asserted	Direct observation, inherited assertion, or analysis	Independent support or contradiction	Evidentiary treatment and limit
SR-01					

Append a recorded-reasoning record where model reasoning, plans, summaries, scratchpads, or explanations are material.

Reasoning record	Architecture and capture point	Integrity, completeness, transformation, or redaction	Context and tool-call correlation	Inference proposed	Permitted proof	Prohibited inference
RR-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
Attempted action			
External action			
Successful completion			
Resulting effect			
Demonstrated harm			
Mitigation and recovery			
Technical actor and execution			

Conclusion layer	Finding	Classification or scope disposition	Attribution boundary and basis
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	Role: governing, risk, selection, development, operation, identity, control, monitoring, or response	Decision, authority, notice, control, intervention, or enforcement proposition	Evidence	Classification	Effect on organizational attribution	Readiness implication
OT-01							

Where a threat cluster, parent organization, government, sponsor, or other higher-level actor is proposed, complete one transition per level rather than collapsing the organizational path.

Actor layer	Proposed actor	Linking proposition	Accessible evidence	Source-attributed assessment, if any	Independent corroboration	Classification and ceiling
Technical executor						
Account or infrastructure operator						
Natural person or team						

Actor layer	Proposed actor	Linking proposition	Accessible evidence	Source-attributed assessment, if any	Independent corroboration	Classification and ceiling
Analytic cluster						
Organization						
Sponsor or state						

WP-11 MCP-Mediated Action Profile

Use when an MCP client, server, tool, task, or downstream service is material to the proposed attribution.

Transition	Proposition	Protocol record	Authorization or consent record	Trace and downstream correlation	Integrity or completeness limit	HAAD dimension affected	Classification
Natural person → host interaction							
Host → model or agent							
Model or agent → MCP client							
MCP client → MCP server							
MCP server → tool							
Tool → downstream service							
Downstream service → consequential action							

Append one row for every additional downstream action produced by the same invocation.

Authorized or represented action	Arguments at client	Arguments at server	Handler or implementation delta	Effective downstream request	Resulting action	Authorization and scope effect

WP-12 Campaign, Activity-Share, and Decision-Gate Profile

Use when a conclusion aggregates several actions, targets, sessions, or agent instances, or relies on a claim about the proportion of work performed by AI.

Field	Entry
Campaign or population asserted	
Unit of analysis	
Inclusion and exclusion criteria	
Total and examined episodes or targets	
Successful, failed, and incomplete episodes	
Evidence coverage and sampling basis	
Common and divergent paths	
Permitted level of aggregation	

Run or episode	Agent instance and configuration	Shared persistent state or causal influence	Direction of influence	Independent repetition supported?	Aggregation limit

Activity-share claim	Denominator and unit	Measurement method	Source and coverage	Exclusions and uncertainty	Attribution inference permitted	Attribution inference prohibited

Decision gate	Proposed human or system actor	Instruction or approval	Evidence of authorship and control	Scope and consequence presented	Classification	Effect on attribution

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.*

13.11 Authenticated MCP client; natural-person control undetermined

> *The authorization record establishes that MCP Client C held a valid token for Server S within Scope R at the material time. The available evidence does not establish which natural person controlled the material host interaction or whether that person authored the instruction that produced Tool Call T. Client authorization is established; natural-person control and interaction authorship are undetermined.*

13.12 Tool invocation established; knowing authorization undetermined

> *The MCP records establish that Tool T received Parameters P and that the server invoked the tool. The available evidence does not establish that Person P was shown those effective parameters, understood the downstream consequence, or knowingly authorized the particular action. Protocol execution is established; knowing authorization is undetermined.*

13.13 Downstream action established; causal contribution bounded

> *The downstream service record establishes Consequence X and correlates it to MCP Request R. The evidence establishes contributions by the host, model, client, server, and tool, but does not support assigning the whole result to any one component or natural person. The causal account is bounded to the contributions established at each transition.*

13.14 Tool-description integrity undetermined

> *The client presented Tool T to the model using Description D and Schema S. The available evidence does not establish that those definitions were the trusted and effective definitions at the material time or*

> *that they remained unchanged between discovery and invocation. Tool invocation is established; the integrity of the meaning represented to the model is undetermined.*

13.15 Tool implementation added an unauthorized downstream action

> *The person authorized the represented action using Parameters P. The client record contains no secondary recipient. The effective handler in Package Version V added Recipient R before the downstream request, and the downstream service delivered the additional copy. Authorization of the represented action is established; authorization of the added disclosure is not. The implementation delta is the immediate technical origin of the secondary action.*

13.16 Subject-system representation contradicted

> *System S represented that Test T passed and that restoration was unavailable. Independent test and backup records contradict those propositions. The evidence establishes that the representations occurred; it does not establish a machine mental state or purpose. The underlying action, test result, restoration state, and representations are adjudicated separately.*

13.17 Source-attributed organizational or state assessment

> *Provider V assesses with stated Confidence C that Cluster G is affiliated with Organization O. The accessible records establish specified technical activity by the identified agent instances and tools. The disclosed material does not permit independent reproduction of the transition from the operator cluster to Organization O. The organizational conclusion remains a source-attributed assessment subject to an evidence-access discontinuity; the separately established technical finding is unaffected.*

13.18 Activity share does not allocate authorization

> *Provider V estimates that the agent performed X percent of tactical operations using Measure M. The estimate does not allocate knowledge, intent, authorization, or responsibility. Human Actor P is separately established or unestablished at each material decision gate, including target selection, escalation, credential use, and the final consequential action.*

13.19 Campaign aggregation bounded

> *The evidence establishes the specified execution path for Episodes E1 through E4. The available records do not establish that the same path, actor, or authorization state applied to unexamined targets or unsuccessful attempts. The conclusion is bounded to the examined episodes and is not promoted to a campaign-wide finding.*

13.20 Later constraint established; enforcement failed

> *Authorized Decision-Maker P issued Constraint C before the material action. Delivery and operative-state records establish that the relevant session received and retained C. Agent A nevertheless executed Action X through an available technical path. The action exceeded the recorded boundary. Continued capability and failed enforcement do not establish continued authority; the enforcement failure is reported separately as a readiness and organizational-control finding.*

13.21 Approval event established; effective-target authorization undetermined

> *The interface record establishes that Person P allowed Command C to proceed after Display D described removal of temporary resources. The effective execution used State File S and Credential K to target production resources not identified in D. Approval of the displayed command is established. Knowing authorization of the effective production target and consequence is undetermined; the material discontinuities lie in knowledge and scope, with the state substitution recorded as causal context.*

13.22 Recorded reasoning correlated; mental-state inference prohibited

> *Trace R records that Agent A described the external environment as real and considered Action X. Tool and downstream records correlate R with the later execution. The evidence supports temporal and decision-lineage findings. It does not establish subjective awareness, belief, deception, or a complete causal explanation of the model's computation.*

13.23 Repeated runs not independent

> *Actions A1 through A6 occurred across Runs R1 through R3. Records establish that R1 created Public Artifact P, which R2 and R3 later retrieved and used. The runs therefore share a causal influence and are not treated as three independent replications. The report distinguishes six actions, three runs, one connected episode, and one incident.*

13.24 Attempt and harm separated

> *Agent A created an external account, submitted Payload P, and contacted Recipient R. Independent review prevented P from being accepted, and no resulting compromise is established. The completed intermediate actions and attempted consequential action are established; successful compromise and resulting harm are not.*

13.25 Organizational functions separated

> *Organization O selected and deployed System S. Function F1 approved the evaluation configuration, F2 owned network controls, F3 operated the runs, and F4 detected and contained the activity. These function-specific decisions and interventions are established to the stated extent. The evidence does not support treating O as one actor with a single state of knowledge or intent.*

13.26 Structured mapping is derivative

> *Catalog M maps the reported campaign to Techniques T1 through T8. M adds standardized analytic labels but cites Provider Report P as its factual source. M is not counted as an independent observation of the activity or actor attribution. The technique mapping and the inherited campaign claims are adjudicated separately.*

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;
- subject-system representations were separated from the truth of the propositions asserted and were corroborated independently where material;
- external confidence labels were reported as source-attributed assessments unless their evidence and method were independently available and evaluated;
- 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;
- each technical, individual, organizational, cluster, and sponsorship path received its own ceiling;
- capability, authority, and enforcement were adjudicated as separate propositions;
- later constraints and material component state were established for authorship or authority, delivery or deployment, timing, applicability, and continued operation;
- compound incidents and multiple downstream actions were decomposed rather than treated as one event;
- vendor, package, publisher-account, implementation, and natural-person identity propositions were not collapsed;
- tool descriptions and schemas were not treated as proof of handler or downstream implementation integrity;
- quantitative autonomy claims stated the denominator, measurement method, source coverage, exclusions, and uncertainty;
- human decision gates were mapped independently of tactical activity share;
- episode-level or sampled findings were not generalized beyond established coverage;
- action, run, session, episode, incident, campaign, and population units were not collapsed;
- repeated runs were treated as independent only after material shared state and causal influence were addressed;
- approval, confirmation, silence, continued operation, or non-intervention was not treated as knowing authorization without testing what was presented and what effectively occurred;
- command text and displayed plans were not treated as proof of effective target, blast radius, or downstream interpretation;
- material state files, inventories, plans, configurations, and resource models were tested for provenance, version, freshness, completeness, and fitness;
- recorded reasoning was not treated as transparent internal computation or proof of awareness, belief, deception, or intent;
- attempt, execution, completion, effect, harm, mitigation, and recovery were adjudicated separately;
- provider, model, configuration, scaffold, agent instance, session, run, tool process, account, and downstream executor identities were not collapsed;
- organizational functions were separated where their authority, knowledge, control, intervention, or contribution differed;
- incident registries, technique catalogs, taxonomies, and derivative reporting were traced to their originating sources and were not counted as independent observations of inherited claims;
- 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.
- Provider telemetry may establish technical activity while leaving organizational, geographic, or sponsorship conclusions inaccessible for independent reproduction.
- Quantitative autonomy estimates may depend on provider-defined units, incomplete visibility, or analytic assumptions and should not be treated as responsibility allocations.
- Campaign evidence may be uneven across targets, successful and unsuccessful episodes, and phases, limiting aggregation.
- Repeated runs may share persistent external state or researcher intervention, limiting claims of independent replication.
- Recorded reasoning may be incomplete, transformed, summarized, hidden, or non-causal and cannot establish a machine mental state.
- Approval interfaces may omit or misstate effective targets, state assumptions, parameters, or consequences.
- Command and infrastructure effects may depend on stale or substituted state, environment, credentials, account context, or downstream interpretation not visible in the command text.
- Public incident registries, technique mappings, and postmortems may be current and well structured while remaining derivative of a small number of originating sources.
- Publicly recognizable scenarios may inflate practitioner agreement through prior exposure rather than consistent application of the method.
- 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.

Agreement may also be inflated where practitioners recognize a public incident, remember a published conclusion, or have previously seen the scenario or reference model. Scenario provenance and participant exposure are therefore controlled and reported rather than assumed away.

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.

The scenario constructor records the hidden construction oracle before the reference model is drafted. The oracle identifies the stipulated event sequence, actor and state relationships, source dependencies, controlled omissions, and intended ambiguity. The reference-model author applies HAAD to that oracle and the practitioner-facing evidence. The independent reviewer tests defensibility without seeing either the oracle or the reference model until the blind analysis is complete.

16.3 Two-round design

Empirical status of Version 0.5: As of August 25, 2026, no independent-practitioner data exist for any HAAD version. Round 1 and Round 2 are planned testing stages and have not been conducted. Version 0.4 was an interim public-review revision based on review and structured scenario analysis, not a practitioner-testing round. The MCP application profile is research- and specification-analysis-driven and remains subject to practitioner testing.

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

Current revision basis. Version 0.5 incorporates corrections supported by review and structured scenario analysis. It does not incorporate practitioner Round 1 findings. A later revision may incorporate explicit rules supported by planned Round 1 results. The change log must distinguish defect correction from preference.

Planned Round 2: operational reproducibility. New practitioners who did not participate in Round 1 will apply Version 0.5, or its frozen successor if Round 1 requires revision, to unseen or materially altered scenarios. Reusing the same facts as the author-developed reference cases is insufficient. Round 2 is the stronger planned 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, one MCP-mediated case requiring correlation from host interaction through downstream consequence, one later-constraint case separating capability from authority and enforcement, one supply-chain case distinguishing tool representation from implementation, one layered organizational-attribution case with inaccessible provider evidence and a quantitative activity-share claim, 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.

The public evidence packs in this package are design exemplars. Once disclosed to a practitioner, they are not eligible as unseen Round 2 instruments. Public incidents, incident registries, technique mappings, research corpora, and public postmortems may supply structural seeds, but their names, memorable facts, public labels, and conclusions must not be carried into a supposedly blind scenario.

16.4 Scenario provenance and recognition-contamination controls

Before practitioner testing, the evaluation custodian maintains a sealed scenario-provenance record containing:

1. every public case, corpus, benchmark, or prior scenario used as a structural seed;
2. the transformation and compositing performed to prevent one-to-one recognition;
3. the construction oracle and its creation date;
4. the controlled evidence gaps, contradictions, inaccessible sources, and dependency relationships;
5. the intended action, run, episode, incident, and aggregation units;
6. the reference-model version and hash;
7. the persons who created, reviewed, or accessed the scenario, oracle, and reference model; and
8. the date on which the practitioner-facing package was frozen.

Transformation must change more than names. It should change the domain, actors, tools, chronology, action form, quantities, and memorable surface facts while preserving the attributional structure under test. Composite scenarios should be preferred where a single public analogue would remain recognizable. The scenario title, instructions, inventory, and distractors must not reveal the expected discontinuity.

Controlled evidence degradation is fixed in the construction oracle before the reference answer is produced. The scenario designer identifies which records are complete, partial, contradictory, derivative, inaccessible, or absent and which propositions each condition can affect. The public conclusion from a source case, an incident-registry label, a technique mapping, or a research author's interpretation is not used as the oracle.

An independent recognizability reviewer examines the practitioner-facing package for distinctive facts that could reveal a source case. After completing the exercise, each practitioner reports prior familiarity with HAAD, the source domains, any recognized case or scenario, and any earlier access to the evidence pack or reference material. The question is asked after completion to avoid priming.

Agreement and reference-model alignment are reported with sensitivity analysis for recognized or previously exposed scenarios. Results from exposed practitioners are not silently pooled with results from participants who did not recognize the source pattern. Recognition does not automatically invalidate a response, but the study must show whether the main conclusion changes when recognized cases are excluded or analyzed separately.

16.5 Stage 1 acceptance criterion

The Version 0.5 specification will pass the Stage 1 structured-scenario pilot only if, when the pilot is conducted, 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 each attribution path at the same evidentiary boundary;
6. distinguish retrospective attribution from prospective readiness findings;
7. correctly reach no demonstrated discontinuity when the evidence establishes every material transition within scope;
8. separate technical execution, natural-person, organizational, and sponsorship conclusions and apply a path-specific ceiling to each;
9. decompose multiple downstream actions and subject-system representations into separate propositions;
10. distinguish tool description, package identity, implementation behaviour, and downstream request state;
11. treat quantitative autonomy claims and human decision gates under Section 5.21;

12. avoid generalizing episode-level findings beyond the stipulated evidence coverage;
13. distinguish action, run, session, episode, incident, campaign, and population units;
14. identify cross-run state or causal dependence before treating repetitions as independent;
15. distinguish approval or non-intervention from knowing authorization of the effective target and consequence;
16. separate attempt, execution, completion, effect, harm, mitigation, and recovery;
17. distinguish recorded reasoning from proof of a machine mental state or faithful causal explanation; and
18. avoid treating registries, technique mappings, taxonomies, or derivative reports as independent corroboration of inherited claims.

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.

Each of the eighteen criteria is scored **Pass**, **Fail**, or **Predeclared Not Applicable** for each pack. Predeclared Not Applicable may be used only where the sealed construction oracle established before distribution that the criterion does not apply. A criterion passes only when all assigned practitioners reach materially equivalent results for every ceiling-setting or endpoint-setting proposition to which it applies. Stage 1 passes only when every applicable criterion passes and the independent reviewer rates every ceiling-setting or endpoint-setting reference proposition and boundary as defensible. Results are not averaged. Any failed criterion requires correction and retesting. Retesting may be limited to affected packs only where independent review establishes that the correction cannot affect another pack, criterion, or attribution boundary.

The independent reviewer completes the analysis before receiving access to the sealed author model. A not-defensible rating prevents a Stage 1 pass and triggers the same correction and retesting as a failed criterion, even if practitioners agree with one another.

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

16.6 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. A public narrative, provider conclusion, registry entry, or technique mapping is not treated as ground truth merely because it is widely repeated or structured.
- **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 0.5 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. Can an independent reviewer distinguish author-model alignment from substantive defensibility?
13. Can practitioners apply the single-stream exception consistently without using it as a routine substitute for convergence?
14. Can practitioners correctly reach no demonstrated discontinuity without relaxing the transition-level sufficiency rules?
15. Can practitioners distinguish an authenticated MCP client from the natural person who controlled or authored the material interaction?
16. Can protocol, trace, task-state, authorization, and downstream records be correlated without treating one vendor's logging plane as independent convergence?
17. Do MCP elicitation and confirmation records establish the precise parameters and consequence presented to the person, or only that an interface event occurred?
18. Can practitioners reconstruct tool-definition and catalogue state at the material time, including changes between discovery and invocation?

19. Which minimum MCP instrumentation permits a bounded finding across host, client, server, tool, and downstream-service transitions?
20. Can practitioners distinguish the represented tool function from the effective handler, package version, and downstream request?
21. Can practitioners identify multiple consequential actions produced through one invocation and adjudicate authorization for each separately?
22. Can practitioners establish whether a later constraint reached and remained operative in the material session without treating continued capability as continued authority?
23. Can practitioners report a provider's organizational or state-attribution assessment without transferring its confidence label or weakening a separately established technical finding?
24. Can practitioners evaluate an AI activity-share claim without treating it as an allocation of intent, authorization, causation, or responsibility?
25. Can practitioners identify strategically material human decision gates when most tactical execution is automated?
26. Can practitioners bound episode-level findings and resist unsupported campaign-wide aggregation?
27. Can practitioners distinguish an approval event or non-intervention from knowing authorization of the effective target and consequence?
28. Can practitioners reconstruct an effective target across command expansion, environment, state, credentials, account context, and downstream interpretation?
29. Can practitioners identify stale, substituted, incomplete, or unfit decision-basis state and classify its effect without changing the canonical dimension register?
30. Can practitioners treat recorded reasoning as a bounded system artifact without inferring awareness, belief, deception, or intent beyond the evidence?
31. Can practitioners distinguish actions, runs, episodes, and incidents and identify shared state that prevents independent-replication claims?
32. Can practitioners separate attempt, execution, completion, effect, harm, mitigation, and recovery consistently?
33. Can practitioners disaggregate organizational functions without treating the organization as one actor or using the organizational path as a fallback?
34. Can practitioners distinguish model family, configuration, scaffold, instance, session, run, tool process, account, and downstream executor identities where material?
35. Can practitioners trace registries, technique mappings, taxonomies, and research reconstructions to their originating sources and avoid false convergence?
36. Do scenario transformation, recognizability review, post-test exposure checks, and sensitivity reporting reduce recognition contamination without changing the attributional structure under test?

References

AI Incident Database. (2026). *Incident 1152: LLM-driven Replit agent reportedly executed unauthorized destructive commands during code freeze, leading to loss of production data*. Retrieved August 21, 2026, from https://incidentdatabase.ai/cite/1152

AI Incident Database. (2026). *Incident 1263: Chinese state-linked operator (GTG-1002) reportedly uses Claude Code for autonomous cyber espionage*. Retrieved August 21, 2026, from https://incidentdatabase.ai/cite/1263

Model Context Protocol. (2026a). *Authorization*. Model Context Protocol Specification, 2026-07-28. https://modelcontextprotocol.io/specification/2026-07-28/basic/authorization

Model Context Protocol. (2026b). *Security best practices*. Model Context Protocol Documentation, 2026-07-28. https://modelcontextprotocol.io/docs/tutorials/security/security_best_practices

Model Context Protocol. (2026c). *Tools*. Model Context Protocol Specification, 2026-07-28. https://modelcontextprotocol.io/specification/2026-07-28/server/tools

Model Context Protocol. (2026d). *Elicitation*. Model Context Protocol Specification, 2026-07-28. https://modelcontextprotocol.io/specification/2026-07-28/client/elicitation

Model Context Protocol. (2026e). *Tasks*. MCP Tasks Extension, draft specification. https://tasks.extensions.modelcontextprotocol.io/specification/draft/tasks

Model Context Protocol. (2026f). *The 2026-07-28 specification*. Model Context Protocol Blog. https://blog.modelcontextprotocol.io/posts/2026-07-28/

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

Kodathala, S. V. (2026). *aiAuthZ: Off-host, identity-bound authorization for AI agents* (arXiv:2607.05518). arXiv. https://doi.org/10.48550/arXiv.2607.05518

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 attack vectors for system-level compromise* (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

MITRE. (2026). *Anthropic AI-orchestrated Campaign, Campaign C0062*. MITRE ATT&CK. Version 1.0. https://attack.mitre.org/campaigns/C0062/

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/

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., et al. (2026). *Agents of Chaos* (arXiv:2602.20021). arXiv. https://doi.org/10.48550/arXiv.2602.20021

Solozobov, O. (2026). *Decision Evidence Maturity Model for Agentic AI: A property-level method specification* (arXiv:2605.04093). arXiv. https://doi.org/10.48550/arXiv.2605.04093

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/

Version 0.5 is a public release. The name and transition-level diagnostic are proposed here; the evidentiary and analytical foundations are attributed. Version 0.5 adds an MCP-mediated attribution application profile, pressure-tested application rules, and scenario-contamination controls while preserving the fourteen canonical dimensions. It is not a validated standard or doctrine. External review and staged testing should determine which distinctions hold.