Artifact-to-Finding Promotion: Evidentiary Requirements for Harm from Correctly Functioning AI Systems

Working Paper, Version 1.1

Kevin V. Watson

August 25, 2026

https://doi.org/10.5281/zenodo.22102862

Correspondence: kevinvwatson@gmail.com

Document information

Author: Kevin V. Watson

Version: Working Paper, Version 1.1

Date: August 25, 2026

Identifier: https://doi.org/10.5281/zenodo.22102862

Correspondence: kevinvwatson@gmail.com

Contents

Abstract

This paper specifies the records required to investigate harm arising from a propositionally accurate AI output whose attributed warrant exceeded its established warrant. It uses artifact-to-finding promotion to name the incident pattern: an AI system performs within specification and a consequential decision assigns its output more evidentiary force than the system established. Existing work separately addresses outcome-graded reliance, contextual interpretation of AI advice, epistemic warrant, and decision provenance. It does not, in the literature reviewed, assemble these elements into an incident-specific reconstruction test for the correct-output case. The paper distinguishes the ordinary technical and domain evidence required to prove input integrity, system operation, and output validity from five promotion-specific records: the output as rendered, qualifiers presented or omitted, the linked decision record, the system's documented warrant at deployment, and contemporaneous model and configuration provenance. It also operationalizes attributed warrant through the documented rationale and observable decision, and distinguishes representation-, decision-, and compound-layer promotion. Worked successful and failed reconstructions demonstrate the method and the consequence of missing records. The central forensic problem is that the decisive promotion record is the output as rendered to the decision-maker, not merely the underlying transaction log, and conventional logging architectures are not necessarily designed to preserve it.

Keywords: AI incidents, digital forensics, AI governance, evidentiary requirements, overreliance, forensic readiness, accountability

Version 1.1 revision note: Version 1.1 preserves the incident category and central argument of Version 1.0 while refining the reconstruction method. It distinguishes foundational validity evidence from the five promotion-specific records, operationalizes attributed warrant through observable decision evidence, adds a layer classification for representation, decision, and compound promotion, and makes the evidence-to-finding pathway, source-independence test, representation-completeness requirement, decision-support boundary, and control-provenance limit explicit. Version 1.0 remains the published record at https://doi.org/10.5281/zenodo.21543521.

1. The Correct-Output Case

Incident taxonomies for AI systems organize around two questions: Did the system fail, and was it attacked?

A third case does not fit either. A system operates as designed, produces an output accurate within its stated scope, logs the interaction completely, and sits at the centre of an event that harmed someone. There is no vulnerability to record, no malfunction to remediate, and no adversary to attribute.

This paper does not claim that territory is unoccupied. It is not. The overreliance literature has developed operational definitions of incidents in which humans deferred to system output without adequate verification, and work on epistemic warrant in AI architectures has established that the

connection between a claim and what makes it true is not preserved by default across generative transformations. Section 3 positions this paper against both.

The claim made here is narrower and rests on two gaps identified in the literature reviewed.

First, dominant operational measures of appropriate reliance are outcome-graded. They generally classify following correct advice as appropriate reliance and following incorrect advice as overreliance. The case addressed here is one where ordinary accuracy checking would find nothing to correct, because the output was accurate. What went wrong was the status assigned to it and the action that status was taken to justify.

Second, adjacent work specifies important components of accountable decision-making, including contextual interpretation, human-review records, system documentation, and decision provenance. It does not assemble those components into an incident-specific reconstruction test for the gap between established and attributed warrant. That synthesis is the paper's principal contribution and the substance of Section 5. The proposed category supplies the object of investigation; the evidentiary specification supplies the method. The latter remains useful even if the former is ultimately treated as a subclass of overreliance.

2. Definition

Artifact-to-finding promotion is harm arising from a decision taken in reliance on the output of an AI system, where the system performed within its specification, the output was propositionally accurate within its stated scope, and the decision assigned more warrant to the output than the system established.

Four elements are load-bearing.

The system performed within specification. If the model was miscalibrated or the training data unrepresentative, the incident implicates established model, data, or specification categories. This is the residual case in which the system did what it was built to do. Operational conformance must be proved from technical evidence; it cannot be inferred merely because no error was reported.

The output was propositionally accurate within its stated scope. A factually wrong output may also be over-trusted, but it does not isolate the failure described here. The defining case is one in which truth at the level of the output does not supply sufficient warrant for the conclusion or action that followed.

A decision followed. This is not a state of belief. It requires an action with consequences. A person who over-trusts an output and acts on nothing has not produced an incident.

The error concerns epistemic status, not propositional accuracy. The output was treated as establishing more than it established. A statement can be true while remaining insufficient to support the decision made from it.

2.1 Artifact and Finding

The distinction the definition rests on is standard in forensic practice, and its absence in AI deployment is the source of the problem.

An artifact is a thing a system produced. Its evidentiary value is that it exists and that the system produced it. It may become evidence for a proposition after its relevance, integrity, completeness, provenance, and limits are tested, but it is not a finding by itself. A finding is a human-adjudicated statement expressing what the available evidence supports, assumes, or cannot resolve within the stated authority, scope, and limits of the inquiry. It carries the reasoning of the person who made it, and that person has to be able to defend it.

In consequential human decision-making, an AI output should ordinarily be treated as an artifact rather than as a finding. A system may be designed and validated to produce determinations autonomously, but that institutional designation does not by itself establish the warrant of a particular output or eliminate the need to preserve how that warrant was derived.

When a detection system flags an account, the flag is an artifact. The conclusion that the account was compromised is an inference, and the inference belongs to an analyst. Collapsing those two things is the same error as treating a file's presence on a device as proof of who put it there. Possession is established. Exclusive control is not. The artifact supports a narrower claim than the one being made.

Promotion, in the sense used here, is the movement of an artifact into the status of a finding without performance of the work that status requires. The incident is what happens when that movement produces a consequence.

2.2 Established and Attributed Warrant

Three quantities must be separated.

Established warrant is what the output actually supports, given how the system was built, what it was designed to detect, and the conditions under which it operated.

Represented warrant is the epistemic force the interface presents the output as carrying. Labels, ranking, emphasis, confidence displays, warnings, omissions, and adjacent context can make the same proposition appear tentative, investigative, or conclusive without changing its content.

Attributed warrant is the epistemic force assigned to the output in the decision process. For reconstruction, it is established through observable evidence: the recorded rationale, propositions asserted, evidence cited, inquiries made or omitted, and action taken. A decision-maker's later account may supplement that record but does not substitute for it. The method does not require an investigator to prove a private mental state.

The incident occurs when attributed warrant exceeds established warrant. Represented warrant is the principal mediating quantity: it can transmit the system's actual limits, or inflate or suppress them before the decision-maker interprets the output. The gap is therefore not a property of the person alone. It may be produced by interface design, qualifiers presented or omitted, procurement and

deployment claims, organizational practice, decision reasoning, or a combination of those conditions. Locating where warrant changed is the analytical task in any investigation of this kind.

2.3 The Specification Boundary

The category requires more than proof that software executed its code correctly. A score is not propositionally accurate merely because the model calculated it without error. Classification requires separate findings about operational performance, the truth or validity of the output within its stated scope, and the warrant attributed to that output.

A practical classification test asks five questions.

1. **Did the system operate as specified?** The deployed model, configuration, inputs, and processing must conform to the documented system state.
2. **Was the output valid within that specification and use context?** Calibration, base rates, threshold selection, population fit, and stated limitations must be adequate for the narrow proposition expressed by the output.
3. **Did represented warrant match established warrant?** The interface's labels, ranking, emphasis, qualifiers, and adjacent context must be compared with the narrow proposition the system validly supported. Inflation at this stage locates responsibility partly in product representation or deployment design; a neutral representation followed by over-attribution locates the gap later in the decision process.
4. **Did the consequential conclusion exceed that narrow proposition?** The decision must depend on an inference the output did not itself establish.
5. **Where did warrant increase?** If represented warrant exceeded established warrant before the decision-maker interpreted the output, classify representation-layer promotion. If representation was proportionate but the documented decision assigned excessive force to the output, classify decision-layer promotion. If both occurred and each materially contributed to the consequential inference, classify compound promotion. A model, data, deployment, or specification failure may coexist, but the pure correct-output category requires investigators to establish that correcting those technical conditions would not eliminate the narrow proposition on which the decision relied.

The categories can coexist. A poorly specified system may produce an output that is also promoted beyond even the warrant the organization claimed for it. Where investigators cannot establish that the output was valid within its stated scope, they should not classify the event as a pure correct-output case. They may record promotion as an additional or provisional classification only if the available evidence independently establishes an unsupported increase in warrant. The residual correct-output category begins only after model, data, calibration, deployment, and specification failures have been considered.

2.4 Promotion-Layer Classification

The category names the incident pattern; the layer classification locates the warrant increase and directs remediation.

Representation-layer promotion occurs where product language, visual emphasis, ranking, omitted qualifiers, deployment materials, or organizational presentation assigns the output more force than its established warrant before the individual decision is made. Remediation principally concerns interface design, deployment claims, and communicated limitations.

Decision-layer promotion occurs where represented warrant is proportionate but the documented reasoning or action assigns the output more force than it supports. Remediation principally concerns adjudication requirements, review practice, decision documentation, escalation, and accountability.

Compound promotion occurs where representation inflated the output and the decision process independently extended it further. The classification should identify both movements rather than force the event into a single layer.

This classification does not convert every interface defect or poor decision into artifact-to-finding promotion. The remaining elements still apply: a consequential decision, reliance on an output valid within its narrow scope, and a demonstrable increase from established to attributed warrant.

3. Related Work and Boundaries

The value of a proposed category depends on whether it does work existing concepts do not. This section states the boundaries against six bodies of work, in order of proximity, and records the scope of the prior-art search.

3.1 Overreliance and Automation Bias

Automation bias is the established term for using automated cues as a heuristic replacement for vigilant information seeking. Parasuraman and Riley (1997) locate overreliance within automation misuse and identify workload, reliability, consistency, and the salience of automation-state indicators as factors shaping it. Empirical cockpit work demonstrated automation bias in consequential decision-support settings (Mosier et al., 1998). Skitka et al. (1999) distinguished omission errors, where an operator misses an event the system did not flag, from commission errors, where an operator follows an automated directive despite contradictory information from more reliable sources. A later systematic review found automation bias across multiple research fields and identified user, system, task, trust, and confidence factors as mediators (Goddard et al., 2012).

More recently, this territory has been operationalized for human-AI decision-making. Schemmer et al. (2022, 2023) define appropriate reliance in terms of distinguishing correct from incorrect AI advice and acting accordingly. On that outcome-graded account, following correct advice is appropriate reliance, rejecting it is underreliance, and following incorrect advice is overreliance. Ibrahim et al. (2025) use a broader formulation that includes accepting incorrect outputs or inappropriately delegating decisions, while their incident classification treats inadequate verification, supervision, or critical evaluation as central.

Those criteria are close to the boundaries proposed here, and the overlap should be stated plainly. Commission-error automation bias is frequently the mechanism by which artifact-to-finding promotion occurs. These are not competing concepts.

Three differences remain.

The output was not wrong. Dominant operational measures are anchored to output correctness. That anchoring is reasonable, because it makes reliance measurable: an answer was correct or incorrect, and the user accepted or rejected it. It also leaves the case addressed here difficult to identify. When the output is accurate and the harm arises from the status assigned to it, checking whether the proposition is true will surface nothing. Investigation must instead test whether the conclusion drawn exceeded the output's scope and warrant. Fok and Weld (2024) identify the limitation of outcome-graded reliance and distinguish it from process-oriented questions about whether reliance was justified.

Unit of analysis. Automation bias is a property of a decision-maker, measured across operators in experimental conditions. Artifact-to-finding promotion is a property of an incident, with a date, a subject, an accountable party, and a record or the absence of one. The relationship resembles that between human error and an aviation accident. One names a mechanism. The other names the thing that gets investigated.

Disciplinary home and remedy. Automation bias sits in human factors and is addressed through training, interface design, and workload management. Artifact-to-finding promotion sits in governance and forensics and is addressed through evidentiary requirements, disclosure obligations, and accountability structures. The literature supports this placement. Skitka et al. (2000) found that making participants accountable reduced automation bias rates, locating part of the remedy in accountability structure rather than cognition.

The closest contextual challenge is Bruijnes et al. (2025). They argue that binary correct/incorrect reliance is inadequate in law enforcement, where predictions are probabilistic, ground truth may be unavailable, and a single output can inform several interventions with different consequences. Their account separates correctness from contextually appropriate action. The present paper accepts that separation and moves it into a different unit of analysis: a harmful incident and the record required to reconstruct it.

3.2 Warrant Preservation in AI Architectures

Romanchuk and Bondar (2026) introduce a Warrant Erosion Principle, holding that epistemic warrant is not automatically preserved through interpreting or generative transformations unless the architecture specifies explicit guarantees preserving the connection between a proposition and its truth-maker. They use **semantic laundering** for the architectural pattern in which propositions with absent or weak warrant acquire epistemic status by crossing trusted interfaces. In a practitioner-facing diagnostic methodology, Kelly (2025) names related patterns including **confidence laundering**, the inflation of uncertainty into unwarranted certainty, and **displaced authority**, the misattribution or erasure of knowledge sources.

This is the closest prior work to the framing in Section 2.2, and the convergence should be acknowledged rather than minimized. Both treat warrant as something that degrades across mediation rather than travelling intact with an output.

Represented warrant is not a renaming of confidence laundering. Kelly's (2025) confidence laundering identifies one pattern that can inflate apparent certainty in an output. Represented warrant is the quantity through which that pattern, semantic laundering, or other interface choices such as ranking, warning language, visual emphasis, and omitted qualifiers mediates between what the system establishes and what a person attributes to it.

The difference is one of layer and purpose. Romanchuk and Bondar (2026) address system architecture and ask what guarantees are required to preserve warrant across tool boundaries. Kelly (2025) supplies an external diagnostic methodology for examining AI knowledge claims before human judgment proceeds. This paper addresses incident classification after a consequential decision and asks what happened, to whom, where warrant increased, and what record would establish it. Architectural guarantees and diagnostic interventions are preventive; the evidentiary requirements here are reconstructive. An organization can adopt either and still be unable to reconstruct a particular incident unless the relevant records were retained.

3.3 Algorithmic Bias and Discriminatory Output

Where a system produces systematically skewed outputs across a protected characteristic, the system has not performed correctly, or the criteria defining correct performance were inadequate. That is a model, data, or specification problem with an established remediation path.

The category proposed here is definitionally the case where correction of an inaccurate output is not the relevant path, because the output was not wrong. The two frequently co-occur, and a biased output subsequently promoted to a finding produces compounded harm. They remain analytically separate because their primary objects of correction differ. Bias may arise in data, modelling, specification, deployment, or institutional practice. Artifact-to-finding promotion is addressed by changing how outputs are represented, interpreted, linked to decisions, and subjected to accountability.

3.4 Terminological Boundaries and Naming History

Two terminological collisions were identified during drafting. Both are recorded here, because the clearance process is itself relevant to a paper arguing for evidentiary discipline.

Interpretive harm. An earlier draft used this as the proposed term. It was withdrawn on discovering established usage in the epistemic injustice literature. Fricker (2007) introduced hermeneutical injustice, occurring when a gap in collective interpretive resources leaves a person unable to render their own significant experience intelligible. Hayes (2024), for example, uses interpretive harm in an analysis of hermeneutical injustice and colonial epistemology. Related work extends epistemic-injustice analysis to AI systems.

That concept and this one run in opposite directions. Hermeneutical injustice concerns a subject who lacks the conceptual resources to understand their own experience. The category proposed here

concerns an institutional actor who possesses ample interpretive resources and applies them to an artifact that does not support the conclusion drawn. The first is a deficit of interpretation. The second is an excess of it. The terms should not be conflated.

Artifact promotion. A subsequent draft used this shortened form. It was also withdrawn. Artifact promotion is established terminology in continuous integration and delivery, denoting the movement of a build artifact between repositories or environments as it passes validation gates. It is implemented as a named feature by major platform vendors and as a named plugin in at least one widely used automation server. Given that the intended readership overlaps substantially with software delivery and platform engineering, the collision was disqualifying.

The compound form retained here carries its own disambiguation. It is longer than a term of art would ideally be. That cost is accepted in exchange for unambiguous reference.

3.5 The Responsibility Gap

Matthias (2004) described a responsibility gap arising where autonomous systems act in ways no human can reasonably foresee or control, leaving no one appropriately held responsible.

The category proposed here is close to the inverse. A person read an output and made a decision, and that person is accountable in the ordinary way. The problem is evidentiary rather than philosophical. The record required to establish what they were shown, what they were told the system established, and what they concluded does not exist. Responsibility is locatable in principle and unprovable in practice.

The practical consequence differs accordingly. Responsibility-gap arguments tend toward restricting deployment or creating new liability regimes. This analysis points somewhere cheaper and more immediate, which is retaining the right records.

3.6 Decision Provenance and Governance Requirements

Decision provenance uses provenance methods to expose the inputs, flows, decisions, and downstream effects within algorithmic systems. Singh et al. (2019) argue that this wider decision pipeline can support oversight, investigation, audit, compliance, and recourse. That work is an important predecessor to the evidentiary requirements proposed here.

Governance instruments supply additional components. The National Institute of Standards and Technology (2023) calls for system knowledge limits, intended uses, human oversight, and contextual interpretation of outputs to be documented. The Treasury Board of Canada Secretariat (2025) requires documentation of decisions and assessments made or assisted by automated systems and access to released software versions for investigation. The Information Commissioner's Office (2026) in the United Kingdom recommends logging human-review decisions and the considerations underlying them. The European Commission (2026) describes logging, traceability, documentation, and human-oversight requirements for high-risk systems under the European Union's AI Act framework.

The distinction is therefore not that relevant records have gone unrecognized. It is that these requirements are distributed across provenance, governance, and human-review frameworks and are

not organized around reconstructing a particular incident in which a true proposition acquired excessive attributed warrant. Section 5 proposes that incident-specific synthesis.

3.7 Search Scope

The literature review was targeted rather than systematic. Searches conducted in July 2026 combined terms concerning correct or accurate AI output, overreliance, automation bias, appropriate reliance, epistemic warrant, artifacts and findings, decision provenance, audit trails, rendered output, human review, and forensic readiness. Sources located through arXiv, ACM, AAAI, PubMed, Springer, Cambridge Core, NIST, the UK Information Commissioner's Office, the European Commission, and the Government of Canada were reviewed. Subscription indexes including Scopus, Web of Science, HeinOnline, and Westlaw were not directly searched. The novelty claim is therefore limited to the literature reviewed rather than asserted universally.

4. Why the Correct-Output Case Warrants Separate Treatment

Four properties, taken together, justify separate investigative treatment.

It is invisible to the security stack. There is no CVE, no anomaly, no policy violation, no signature. Access control was honoured. Authentication succeeded. The logs are clean and accurate. Every control performed as designed. No monitoring system in a conventional enterprise environment will indicate that the incident occurred.

It has no remediation path in the conventional sense. There is no patch, because nothing was broken. The correction applies to the decision process surrounding the system, which typically sits outside the boundary security functions are resourced to work within. This creates an ownership vacuum. The people who can see the system cannot change the decision process, and the people who own the decision process cannot see the system.

It produces harm falling on a party outside the interaction. Adversarial compromise harms the organization operating the system. Artifact-to-finding promotion typically harms whoever the output was about. An employee flagged by a monitoring tool. A claimant scored by a model. A person named in an automated report that was read as an accusation. That asymmetry places the category in governance rather than usability. The organization holds the interpretive authority. The person bearing the consequence has no visibility into how the output was produced, what it supported, or that it was involved at all.

It is systematically under-recorded. This follows from the first property and compounds the others. Because no control was tripped, no incident is opened. Because no incident is opened, no post-incident review occurs, no entry appears in incident statistics, and the organizational learning loop never closes. An organization can sustain these incidents repeatedly and hold a complete, accurate incident record showing none.

The fourth property is the argument for a named investigative pattern rather than improved practice alone. Practices are refined by feedback. A failure mode generating no feedback does not get refined. Naming it is the minimum condition for counting it.

5. Evidentiary Requirements

An incident pattern is only useful if events within it can be investigated. This section synthesizes elements that existing work treats separately and applies them to reconstruction of the correct-output case.

The reconstruction question is not limited to what the system did. It asks what evidentiary force the output was presented as carrying, what force the documented decision assigned to it, whether that assignment was supported by what the system established, and where any increase entered the chain.

A complete reconstruction has two evidentiary parts. The first is **foundational validity evidence**: the inputs and source records relevant to the particular output; evidence of input integrity and processing; applicable validation, calibration, population-fit, threshold, and limitation evidence; and any domain evidence needed to test the narrow proposition. This material establishes whether the system operated as specified and whether the output was valid in context. It is not unique to artifact-to-finding promotion, but the correct-output classification cannot be made without it.

Source material, evidence, analytical output, and finding are not interchangeable. A log entry, file artifact, AI output, vendor statement, prior report, dashboard, export, or tool result begins as source material. It supports a finding only after the relevant proposition has been stated, the material has been tested for the question, competing explanations have been considered, and a person has adjudicated what the record supports and what remains unresolved.

The second part is a package of **five promotion-specific records**. They do not all perform the same evidentiary function. The rendered output, including a distinguishable record of qualifiers presented or omitted, and its linkage to the consequential decision are load-bearing for classification: without presentation and linkage, represented and attributed warrant cannot be tied to the harm. Deployment-time warrant and model/configuration provenance establish the baseline against which promotion is measured and enable competing model, specification, and deployment explanations to be tested. Provenance identifies the system state; it does not replace the input, validation, or domain evidence needed to prove that a particular output was valid.

The output as rendered. Not the underlying data alone, but what the person saw: content, ranking, emphasis, adjacent context, and visual weight. Where the material presented to a human may differ from the machine-readable source or an intermediate transformed representation, preserve and compare the relevant representations and document any material difference. Different views of the same source are not independent corroboration merely because they look different. This is the primary incident artifact, yet transaction logs designed around system operations may not retain it.

Qualifiers presented or omitted. Confidence values, scope statements, known limitations, error rates. The omission of a qualifier is itself an artifact of the incident and must be capturable as such. A record

that cannot distinguish "no confidence value was displayed" from "no confidence value was retained" is inadequate.

The decision record, linked to the output. The action taken, by whom, its temporal relationship to the output, and the observable basis assigned to the action: the rationale recorded, propositions asserted, evidence cited, review performed, alternatives considered, and material inquiries made or omitted. Linkage matters more than either record alone. Two accurate records with no established relationship will not support a finding.

The system's documented warrant at deployment. What the system was designed to establish, recorded when it was deployed rather than reconstructed afterward. Post-incident reconstruction is subject to obvious pressure and will not withstand scrutiny.

Model version and configuration provenance at the time of output. Without it, the output cannot be tied to a system state. This record permits the applicable specification and validation evidence to be identified; it does not by itself verify the truth or validity of the output.

Where attribution or accountability is material, the investigator also requires proportionate **control provenance**: evidence of who or what designed, configured, authorized, deployed, operated, modified, supervised, reviewed, or intervened in the system and decision process. Control provenance is not a sixth classification-critical promotion record. It addresses a different question: who or what exercised material control over the conditions in which warrant was represented and assigned. An account, credential, approval record, administrative event, or organizational role does not by itself establish human identity, authorship, knowledge, authorization, intent, causation, or responsibility.

5.1 Evidence-to-Finding Reconstruction Pathway

The reconstruction follows a controlled pathway:

Pathway: Source material -> Observation -> Claim -> Tested claim -> Human-adjudicated finding -> Decision support

The pathway is not a confidence ladder. Each transition changes the status of the proposition and requires a stated basis.

1. **Set authority, scope, and the consequential decision.** Identify the decision or action under examination, the affected party, the relevant time period, and the authority and limits of the inquiry.
2. **Preserve the material record.** Preserve the output and its relevant visible, machine-readable, and transformed representations; qualifiers; decision linkage; deployment documentation; model and configuration provenance; and, where attribution is material, control provenance.
3. **Establish the narrow technical and domain proposition.** Test system operation, input integrity, processing, validation, calibration, population fit, thresholds, and material limitations. A specification, vendor statement, prompt, policy, or deployment description is a claim about the intended condition, not proof that the deployed condition existed.

4. **Map the warrant transitions.** State the established warrant, identify the represented warrant, reconstruct the attributed warrant from observable decision evidence, and locate any increase at the representation layer, decision layer, or both.
5. **Test competing explanations and independence.** Consider model, data, specification, deployment, interface, decision-process, and unrelated explanations capable of defeating or changing the classification. Corroboration is independent only to the extent that it arises through a materially separate process, control point, system layer, actor, or observation path. Repeated dashboards, summaries, or exports from the same evidentiary pipeline do not corroborate one another.
6. **Adjudicate and report.** A human investigator states what is established, what is assumed for a limited purpose, and what cannot be resolved. Each finding identifies its evidence, reasoning, material contrary evidence, assumptions, and limits. The result supports the authorized decision-maker. It does not make the organizational, legal, employment, operational, or governance decision.

This pathway also prevents a prior report or system-generated explanation from inheriting finding status. Either may identify a lead, claim, or gap. Neither corroborates itself, and neither substitutes for inspection of its underlying evidence and reasoning.

5.2 The Central Forensic Problem

The first record is the difficult one, and the difficulty is architectural rather than incidental.

The decisive record in an artifact-to-finding promotion incident is the output as rendered, not merely the underlying data or transaction log.

Logging architectures capture data operations. Queries, transactions, authentication events, API calls. They do not capture presentation. Yet represented warrant enters at the presentation layer, because that is where a human encountered the output before assigning it a role in a decision. An organization can hold complete, accurate, tamper-evident logs and be unable to answer the first promotion-specific question, which is what the person was actually shown.

This is a specific instance of a general design risk. Audit logging may be implemented at the layer where the technology sits rather than the layer where the decision happens. When those layers differ, the audit trail can be complete, accurate, and unable to address the event. Nothing is missing from the logs. The logs were never scoped to the question.

Architecture is not the only explanation. Retaining presentation-layer records creates an organizational tradeoff. The record supports accountability and recourse, but it also increases storage and security obligations, may retain personal information beyond an operational need, and creates evidence available to access requests, regulators, discovery, or litigation. Canadian privacy guidance illustrates both sides: personal information should generally be retained only as long as necessary, yet information used to make a decision about a person should be retained long enough to permit access and challenge (Office of the Privacy Commissioner of Canada, 2025). The benefits of retaining a rendering often fall to the decision subject and a future investigator, while its immediate cost and exposure fall to the

organization deciding whether to build the capability. That distribution creates a structural incentive toward under-retention even without deliberate concealment.

This incentive must not be confused with spoliation. A failure to design a system to create a record is analytically different from destroying or failing to preserve relevant evidence once litigation, an access request, an investigation, or another preservation duty has arisen. Legal hold changes the organization's position: ordinary disposal processes may have to be suspended for existing relevant records, and litigation readiness depends on the ability to preserve and produce electronic records (Department of Justice Canada, 2018). Spoliation rules vary by jurisdiction and generally require facts beyond mere non-creation. The governance problem arises earlier. If the decisive rendering was never made durable, a later preservation duty cannot recover it.

The practical test is inexpensive and should be applied before deployment. State the investigative question first. Then ask which layer holds the record that answers it. That test is cheap to run in advance and expensive to skip.

5.3 Worked Reconstruction

Consider a hypothetical employee-monitoring incident. An identity-security system generates an impossible-travel alert after recording successful access to an employee's account from Winnipeg and Amsterdam eighteen minutes apart. The underlying authentication events and IP geolocation results are accurate, the rule fires as configured, and the system is documented as detecting geographically inconsistent access. A security manager sees the alert and concludes that the employee shared credentials with an unauthorized person. The employee is dismissed.

The system established a narrow proposition: the account produced two successful access events associated with geographically distant IP locations inside a time interval inconsistent with physical travel. It did not establish who performed either access, whether the geolocation reflected physical presence, whether a corporate VPN or cloud service affected the apparent location, or whether the employee acted improperly. The dismissal rests on those additional inferences.

The foundational validity evidence and five promotion-specific records permit the incident to be reconstructed.

Foundational validity evidence establishes the technical and factual predicate. The retained authentication records establish the two successful account events; input-integrity evidence establishes that the event data supplied to the rule were not altered; the applicable geolocation documentation and deployment validation establish the supported use and known limitations of the location inference; and the configured rule can be tested against those inputs. Together with provenance, this evidence supports the narrow conclusion that the rule operated as specified and that the alert was valid for triage. It does not establish the human actor, physical presence, credential sharing, intent, or misconduct.

The output as rendered establishes what entered the manager's decision environment. Suppose the retained rendering shows a red "Critical: Impossible Travel" banner, the employee's name, a map joining Winnipeg and Amsterdam, and the alert ranked first in the review queue. It also establishes that the

interface did not merely display two authentication events; it represented their warrant through visual and linguistic cues carrying urgency and apparent conclusiveness.

Qualifiers presented or omitted establish the limits communicated at the moment of interpretation. Suppose the rendering contains no visible statement that IP geolocation is inferential, that VPN routing can produce geographically inconsistent locations, or that the alert does not identify the human actor. A limitations page elsewhere in the product does not establish that those qualifiers were available to this manager at this decision.

The linked decision record establishes the consequence and its relationship to the output. The dismissal record identifies the manager, time, action, and stated reason: credential sharing inferred from the impossible-travel alert. A shared incident identifier links that record to the rendered alert rather than merely showing that an alert and a dismissal occurred on the same day.

The system's documented warrant at deployment establishes the proposition the organization was entitled to draw from the system. The deployment record describes the rule as a triage signal for geographically inconsistent account access requiring investigation. It does not describe the rule as establishing credential sharing, employee identity, intent, or misconduct.

Model version and configuration provenance establishes that the alert can be tied to the system state under examination. It identifies the geolocation database, rule threshold, allowlists, VPN exceptions, software version, and configuration active at the time. Those records confirm that the system performed as documented and permit investigators to test whether an excluded configuration or known defect offers a competing explanation.

On those records, an investigator can make four separate findings. The alert was operationally and propositionally valid within its narrow scope. The organization had documented that its warrant stopped at triage. The rendered interface increased that warrant through apparent conclusiveness and omitted limitations. The dismissal record then assigned the alert force extending to identity and misconduct without documenting independent support. The event is therefore compound promotion: represented warrant exceeded established warrant, and the decision extended the unsupported inference into a disciplinary finding. The harm does not depend on showing that the alert was false. This forensic classification supports review of the dismissal; it does not itself decide the employment outcome or assign legal responsibility.

The reconstruction could also defeat the proposed classification. If provenance showed that the geolocation database was outside its supported region, that the rule was uncalibrated for the deployed population, or that the organization had specified the alert itself as proof of credential sharing, the investigation would move toward model, deployment, or specification failure. The method is not designed to force events into a new category. It is designed to locate the failure at the layer the evidence supports.

The worked case also exhibits the four properties in Section 4. No security control failed, patching the alert would not correct the unsupported attribution, the employee bore the consequence outside the

system interaction, and the event would not enter an incident register unless the organization recognized the interpretive failure.

5.4 Failed Reconstruction

Now remove two records from the same case. The system retains the authentication events, the alert rule, and the model configuration, but not the rendered screen. The dismissal record states only "security policy violation" and carries no identifier linking it to the alert.

An investigator can establish that the account generated geographically inconsistent access and that the impossible-travel rule fired correctly. The investigator cannot establish whether the manager saw a critical banner or a neutral triage notice, whether limitations were displayed, whether the manager relied on the alert, or whether credential sharing was inferred from separate evidence. The system record and employment record are individually authentic but their relationship is unproved.

The result is not a finding that artifact-to-finding promotion occurred. It is an undetermined result: the available record cannot distinguish promotion from correct triage followed by an independently supported decision. Foundational validity evidence and deployment documentation may establish the alert's narrow warrant, but represented and attributed warrant cannot be reconstructed. The missing rendering prevents analysis of what epistemic force the interface conveyed; the unlinked decision record prevents analysis of what force the documented decision assigned to it. The investigator should state whether resolution remains pending because identifiable evidence or work is outstanding, or exhausted because available and obtainable evidence has been pursued to the proportionate limit of the inquiry.

That inability is the consequence predicted by the fourth property. No technical incident was opened, so the presentation and decision linkage were not preserved. A complete transaction log remains incapable of answering the investigative question. The failed reconstruction therefore distinguishes evidentiary roles within the promotion-specific package. The rendered output, its qualifier state, and its link to the consequence are classification-critical; omission of presentation or linkage can make the incident unclassifiable. The deployment-warrant and provenance records establish the authorized baseline and system state. Their absence reduces the strength and specificity of the available finding and may prevent investigators from excluding specification or deployment failure.

6. Limits of the Claim

Several limits should be stated.

The category is proposed, not validated. It rests on analytical argument and on the structure of existing criteria. It is not supported here by a case series and should be treated as a hypothesis for testing against casework rather than an established finding.

The test in Section 2.3 makes the specification boundary explicit and structures the dispute, but it does not make the boundary mechanical. Whether calibration, thresholds, population fit, and deployment conditions make an output valid within its stated scope requires domain judgment and may not yield

agreement between assessors. The specification is also an organizational claim rather than a neutral fact. Where a vendor or deployer describes outputs as findings, determinations, or conclusions, the apparent promotion may reveal an inadequate specification, representation-layer promotion, or both. The reconstruction method identifies the propositions, layers, and records over which the parties disagree; resolving them still requires domain evidence.

The distinction between artifact-to-finding promotion and inadequate specification requires judgment that will vary between assessors. Reducing that variance is a matter for method development and is not attempted here.

The literature review supporting Section 3 was targeted rather than systematic and did not directly search several subscription indexes. Additional prior art may exist, particularly in medical decision support, aviation, administrative law, and legal evidence scholarship, where contextual reliance and the artifact-and-finding distinction have long histories.

6.1 Validation Agenda

The next step is not to accumulate examples that fit the category, but to test whether independent investigators can apply the reconstruction method consistently and whether it changes what an inquiry can establish.

A useful first study would assemble a purposive case set containing four kinds of events: incorrect outputs followed by harm, correct outputs appropriately used, correct outputs promoted beyond their warrant, and cases made indeterminate by missing records. Cases should be drawn from more than one consequential domain and should include negative cases capable of defeating the proposed classification. Investigators who did not develop the framework would receive the five-question test and the available record package, but not the intended classification. They would separately classify system performance, propositional validity, represented warrant, attributed warrant, the locus of any warrant increase, and whether the final classification was determinate.

The primary measures should be inter-rater agreement for each classification question, the proportion of cases remaining indeterminate, and the reasons for disagreement. A second comparison should test the same cases with and without the rendered-output and decision-linkage records. If those records are genuinely classification-critical, their removal should predictably reduce agreement or convert otherwise classifiable cases into indeterminate ones. If investigators reach the same conclusions without them, the five-record specification is overstated and should be revised.

The framework should also be tested prospectively as a forensic-readiness control. Organizations can map both the foundational validity evidence and the five promotion-specific records to an existing AI-assisted decision process, run a simulated reconstruction, and measure whether the records can be retrieved, linked, and interpreted within a defined period. This would test operational feasibility, storage and privacy costs, and whether presentation capture can be made sufficiently stable to support later inquiry.

Three results would count against the present claim: investigators cannot distinguish the category reliably from automation bias or inadequate specification; represented warrant cannot be assessed with

acceptable consistency; or the proposed records do not materially improve reconstruction. Stating those failure conditions matters because the category should survive by discriminating cases and improving findings, not merely by redescribing harm after it occurs.

7. Conclusion

Organizations deploying AI systems into decisions about people are building an evidentiary position, whether or not they intend to. That position may be weaker than operational logging suggests. The logs can be complete. The controls can be functioning. Yet an organization may still be unable to establish how a specific decision affecting a specific person was reached and what the system's output actually supported.

Artifact-to-finding promotion names the case that exposes this. The event is invisible to security controls, unremediable by patching, borne by someone outside the interaction, and absent from the incident record. Those four properties describe a failure mode that will not be discovered by the mechanisms currently in place to discover failure modes.

The practical contribution is the reconstruction test. Foundational validity evidence establishes what happened in the system and whether the narrow output was valid in context. Five promotion-specific records establish the output as rendered, the qualifiers presented or omitted, the decision linked to that output, the system's deployment-time warrant, and contemporaneous model and configuration provenance. Together, the two evidentiary parts permit an investigator to separate system operation, propositionally valid output, represented warrant, documented decision rationale, and the layer at which unsupported authority entered the decision.

Authority comes from method, not assertion. Evidence sets the boundary of what can be claimed. Applied to AI-assisted decisions, that principle has a specific consequence. Without the record, there is no boundary, and anything can be claimed. Including, most consequentially, that the system found it.

References

Bruijnes, M., Grimmelikhuijsen, S., & Robeer, M. (2025). Explainable AI is no silver bullet: Towards a contextual understanding of appropriate reliance on AI in law enforcement. In J. Goossens, E. Keymolen, & A. Stanojević (Eds.), *Public governance and emerging technologies*. Springer. https://doi.org/10.1007/978-3-031-84748-6_4

Department of Justice Canada. (2018). *Audit of e-discovery and litigation readiness*. https://www.justice.gc.ca/eng/rp-pr/cp-pm/aud-ver/2018/edis/p1.html

European Commission. (2026). *AI Act*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fok, R., & Weld, D. S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine, 45*(4), 494–513. https://doi.org/10.1002/aaai.12182

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hayes, C. (2024). Hermeneutical injustice via interpretive harm: Epistemologies of change for structural oppression in Africa. In N. Tshishonga & I. Tshabangu (Eds.), *Democratization of Africa and its impact on the global economy* (pp. 18–31). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0477-8.ch002

Ibrahim, L., Collins, K. M., Kim, S. S. Y., Reuel, A., Lamparth, M., Feng, K., Ahmad, L., Soni, P., El Kattan, A., Stein, M., Swaroop, S., Sucholutsky, I., Strait, A., Liao, Q. V., & Bhatt, U. (2025). *Measuring and mitigating overreliance is necessary for building human-compatible AI*. arXiv. https://doi.org/10.48550/arXiv.2509.08010

Information Commissioner's Office. (2026). *Human review*. https://ico.org.uk/for-organisations/advice-and-services/audits/data-protection-audit-framework/toolkits/artificial-intelligence/human-review/

Kelly, M. (2025). *The epistemic suite: A post-foundational diagnostic methodology for assessing AI knowledge claims*. arXiv. https://doi.org/10.48550/arXiv.2510.24721

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology, 8*(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Office of the Privacy Commissioner of Canada. (2025). *PIPEDA fair information principles*. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Romanchuk, O., & Bondar, R. (2026). *Semantic laundering in AI agent architectures: Why tool boundaries do not confer epistemic warrant*. arXiv. https://doi.org/10.48550/arXiv.2601.08333

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). *Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making*. arXiv. https://doi.org/10.48550/arXiv.2204.06916

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery. https://doi.org/10.1145/3581641.3584066

Singh, J., Cobbe, J., & Norval, C. (2019). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access, 7*, 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.

Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*(4), 701–717.

Treasury Board of Canada Secretariat. (2025). *Directive on automated decision-making*. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Author Note

Kevin V. Watson is a Digital Forensics and Incident Response specialist and a graduate student in the Master of Interdisciplinary Artificial Intelligence program at the University of Ottawa. He is the author of the Zemi Method, a published doctrine on investigative reasoning and evidentiary boundaries, available at https://zemimethod.com. Correspondence: kevinvwatson@gmail.com.

This paper is issued as a working paper. The category it proposes is offered for testing and criticism rather than as a settled result. Correspondence and challenge are welcome.

Suggested citation: Watson, K. V. (2026). *Artifact-to-finding promotion: Evidentiary requirements for harm from correctly functioning AI systems* (Working Paper v1.1). Zenodo. https://doi.org/10.5281/zenodo.22102862.