

emanuel
celano
19.09.2026
09:48:09
GMT+02:00



EVIDE

External Evidentiary Deposit
<https://app.certifywebcontent.com>

**PROTECT INNOVATION. PRESERVE EVIDENCE. KEEP JUDGMENT
WHERE AUTHORITY BELONGS.**

Building Better AI without Losing the Evidence

*Preserving What Matters for AI Providers, Users and
Independent Examination*

Evidence Before Interpretation

Better AI needs better evidence, not more exposure.

EVIDE Research Discussion Note

A research discussion note, shared for external comment. Not a legislative proposal or a technical specification.

Dr. Emanuel Celano
Founder & Chief Architect, EVIDE
September 2026

Status and scope of this note

This note is a working synthesis of a multi-round, adversarial research exercise conducted against the EVIDE hypothesis described in Sections 1 to 6 below. The exercise deliberately tried to falsify the claim that the hypothesis identifies something new, by searching EU law, technical standards, and the law and regulatory practice of the United States, the United Kingdom, South Korea and China for functional equivalents, whatever their terminology.

The findings are provisional. They are well grounded for EU law (AI Act, Machinery Regulation, Product Liability Directive, Cyber Resilience Act, General Product Safety Regulation) and reasonably grounded for the United States, the United Kingdom, South Korea and China. They are not yet grounded to the same standard for Canada, Japan, IEC 61508, ISO 26262, the IEEE standards portfolio, the NIST AI Risk Management Framework, or the Medical Device Regulation. Section 9 lists these gaps explicitly. No claim in this note should be read as final until those sources are checked.

Terminology note. An earlier working name for the observable phenomenon at the centre of this note, "Decision Conflict Signature", has been abandoned. It borrowed a causal vocabulary from cognitive psychology that the underlying observation cannot support: a statistical or arbitration-level deviation does not establish that a system experienced anything resembling conflict. The current working

name is Decision Boundary Deviation (DBD). This note does not claim that a robot's internal state can or should be inferred from a DBD.

The idea in simple terms

Imagine an autonomous robot performing its normal tasks.

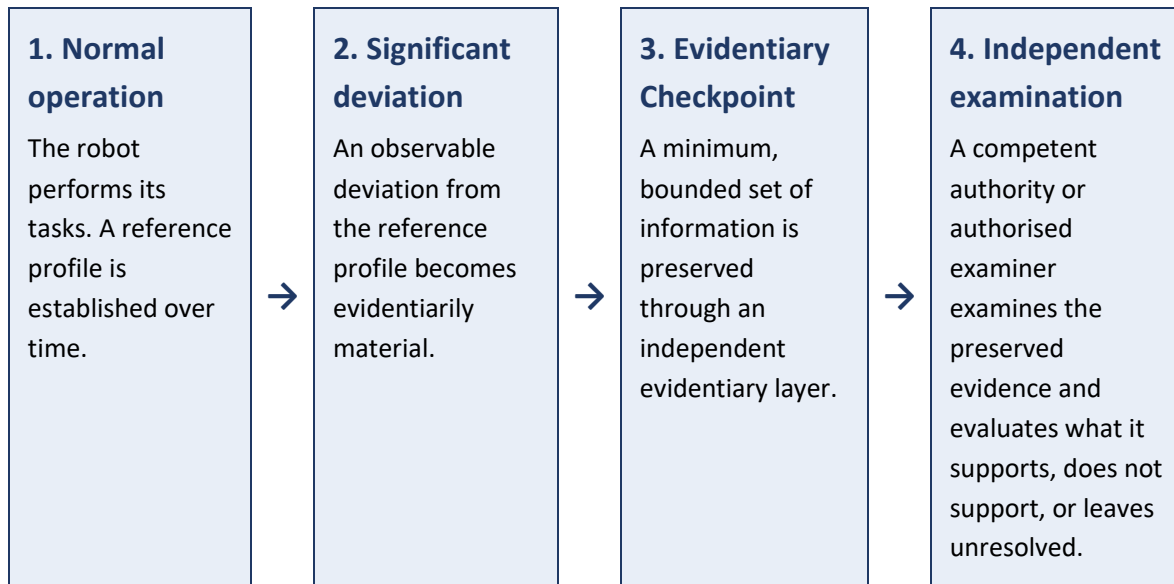
Over time, certain observable aspects of how the system approaches a decision can form a reference profile: for example, how long particular decision phases normally take, whether certain policy or rule checks are consulted, how often alternatives are reconsidered, or other externally measurable signals appropriate to that system.

This paper asks a simple question: what if a significant deviation from that reference profile could trigger the preservation of evidence before anyone knows whether something has actually gone wrong?

The deviation itself would prove nothing. It would not mean that the robot made a mistake, violated a rule, became unsafe, or experienced an internal "conflict". It would simply mean: this moment is sufficiently unusual to preserve a limited record of what happened around it.

The paper calls the reference mechanism an Evidentiary Metronome. When a predefined and measurable deviation becomes evidentially material, it can activate an Evidentiary Checkpoint. The checkpoint preserves only a minimum,

bounded set of information needed to make later examination possible.



**The deviation does not accuse the robot.
It triggers the preservation of evidence.**

What can the preserved evidence help us understand?

The purpose of preserving evidence is not simply to prove that a deviation occurred. It is to make later questions easier to examine independently. Depending on the Minimum Evidentiary Surface available for a given system, a preserved checkpoint could help a competent authority or an authorised examiner ask questions such as:

- What was the system actually observing at the relevant moment?
- Which model, software, policy and safety configuration was in force?
- What observable behaviour departed from the declared reference profile, and by how much?
- What happened immediately before and after that departure?

- Did a safety mechanism, override, fallback or other control subsequently intervene?
- Does the preserved record support the account later provided by the system or provider, or does it leave a discrepancy unresolved?
- Is there enough independently preserved material to reconstruct the relevant sequence without relying exclusively on the manufacturer's own records?

EVIDE does not decide what happened. It preserves the evidentiary material that may allow a competent authority or authorised examiner to evaluate what happened.

The purpose of independent preservation is therefore not to replace the authority, investigator or examiner, but to give them a bounded, integrity-preserved record that can become material for their own subsequent examination.

The evidence may not answer every question, and it does not determine fault, liability or correctness. Its value is that these questions can be asked later against a record preserved at the relevant moment, rather than only against explanations reconstructed after the event.

The question this section answers, in short, is not what the AI was thinking, but what can still be independently examined after the event.

But there is another essential part of the proposal: the evidentiary record should not remain solely under the control of the company whose system may later need to be examined.

A robot manufacturer may already maintain extensive internal logs, telemetry and diagnostic information. Those records can be extremely valuable. But evidence intended to support independent post-event examination raises a different requirement: preservation should be structurally separated from the sole control of the system or provider being examined.

This does not mean exposing the manufacturer's intellectual property. The proposed evidentiary surface can remain deliberately narrow. Model weights, source code, chain-of-thought, proprietary reasoning and unnecessary internal states do not need to be transferred merely because a checkpoint is triggered.

Instead, the checkpoint can preserve a minimum evidentiary set through an independent evidentiary layer, with integrity, provenance and version-binding information sufficient for later examination.

The distinction is fundamental: the manufacturer continues to build and improve the AI. The evidentiary layer preserves the bounded record. The competent authority or authorised examiner evaluates what that record supports, does not support, or leaves unresolved.

No part of that sequence requires the evidentiary trigger itself to decide whether the robot was right or wrong. Its purpose is narrower: preserve enough evidence, at the right moment, outside the sole control of the party whose conduct that evidence might later help examine.

In one breath: the robot deviates -> the deviation does not accuse it -> an independent layer preserves the minimum necessary -> someone else, with the authority to do so, evaluates afterwards what the evidence supports, does not support, or leaves unresolved.

Contents

The idea in simple terms

1. The Problem
 2. The Missing Temporal Layer
 3. The Proposed Distinction
 4. The Evidentiary Metronome
 5. Minimum Evidentiary Surface
 6. The Confident-Wrong Limitation
 7. Existing Regulatory and Technical Components
 8. The Compositional Gap
 9. Open Prior-Art Dependencies
 10. Research Question
- Appendix: Element-by-Element Summary (A-K)

1. The Problem

Three regulatory functions currently do most of the work around autonomous and AI-enabled systems: the technical capability to log events continuously over a system's lifetime, the recording of interventions once a safety mechanism has already acted, and the reporting of an incident once harm, or a reasonable likelihood of harm, has already occurred.

These three functions are well established and, in the European Union in particular, increasingly binding. Article 12 of Regulation (EU) 2024/1689 (the AI Act) requires high-risk AI systems to technically allow automatic logging over their lifetime. Regulation (EU) 2023/1230 (the Machinery Regulation) requires a tracing log tied to safety interventions, together with binding of the software version in force at the time. Article 73 of the AI Act requires providers to report serious incidents once a causal link, or a reasonable likelihood of one, has been established.

None of these three functions asks a fourth question: is there a moment, prior to any determination that harm occurred or that a safety mechanism had to intervene, at which an autonomous system's behaviour already departs from what would ordinarily be expected of it, in a way specific enough to be worth preserving before anyone knows what it means?

This note treats that as an open, and so far unanswered, question, and reports on an attempt to find out whether the question has already been answered elsewhere under different terminology.

2. The Missing Temporal Layer

The research distinguishes three categories of trigger, used consistently throughout the technical and legal review that supports this note.

A. Outcome trigger

Something serious has already happened. The system reports or records because harm, a crash, a death, or a comparable outcome has occurred or is reasonably likely to have occurred. This is the logic of Article 73 of the AI Act, of crash-triggered event data recorders, and of accident notification under the General Product Safety Regulation.

B. System-state trigger

A fault, an override, or a safety intervention has already occurred inside the system. The trigger fires because a distinct, already-defined mechanism, a controller switch, a safety filter override, a fallback, has already acted. This is the logic of the Machinery Regulation's tracing log and of runtime assurance architectures such as Simplex, where control authority moves from an unverified controller to a verified backup.

C. Evidentiary materiality trigger

Something departs, to a degree declared in advance, from a declared reference profile. The trigger fires before anyone, including the system itself, has established what the departure means. It does not presuppose harm, fault, or a safety intervention. It only marks the moment as a candidate for preservation.

The research summarised in Section 7 found extensive coverage of categories A and B across EU law, US federal regulation, and Chinese national standards. It found no verified instrument, in any of the jurisdictions checked so far, that imposes or standardises category C.

3. The Proposed Distinction

The practical value of separating category C from A and B is temporal, not evaluative. It changes what the deviation is asked to do.

Deviation -> correct the robot
(a control-engineering response, not
evidentiary)

Deviation -> declare an incident
(an outcome-trigger response, arrives
too late)

Deviation -> the robot made an error
(a judgment the deviation cannot
support)

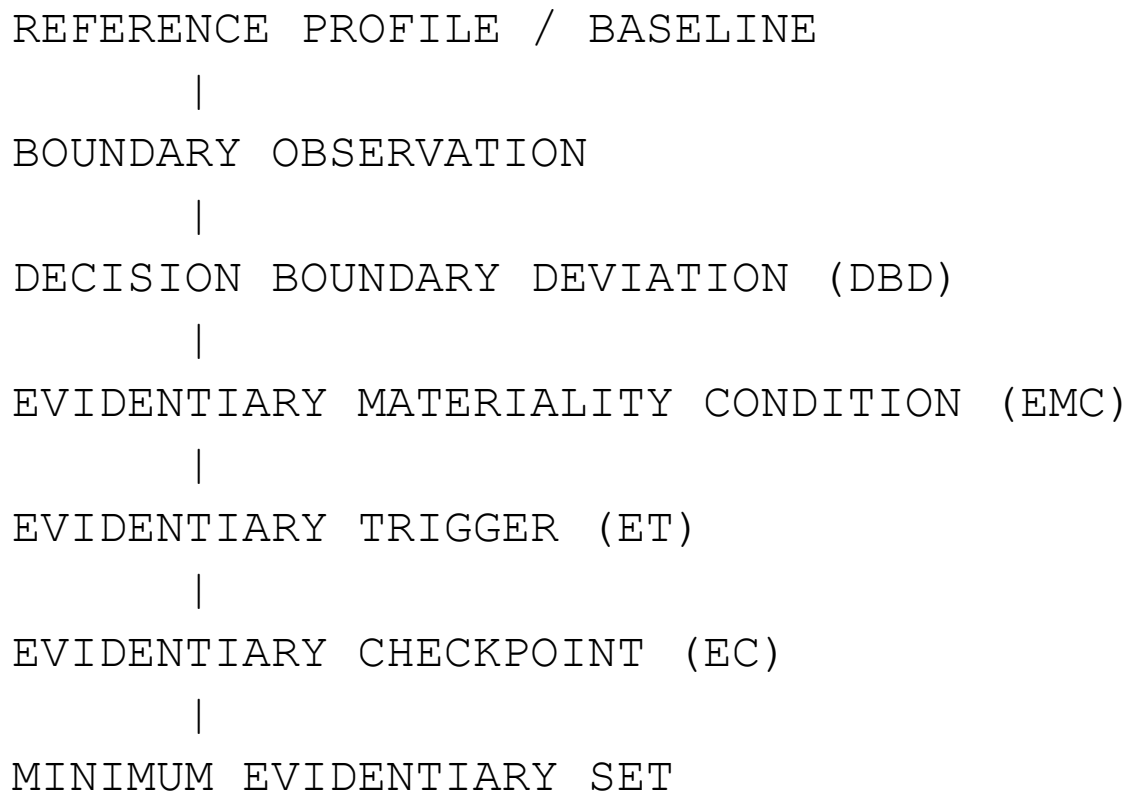
The proposed alternative reads instead:

Deviation -> this moment may later
matter -> preserve it before anyone
knows why

The distinction matters because a category C trigger, if it is to be defensible in an evidentiary setting, must be constructed so that firing it never amounts to a finding of error, fault, danger, or internal conflict. A deviation from a declared reference profile is a fact about observable behaviour relative to a declared baseline. It is not, by itself, a fact about causation, responsibility, or safety. Whether it turns out to matter is a question for later, independent examination, not for the mechanism that decided to preserve it.

4. The Evidentiary Metronome

The working architecture proposed for a category C mechanism, informally called the Evidentiary Metronome, chains seven stages:



The reference profile is not a simple average. Technical review found that a single baseline per decision class is likely to be statistically weak, because real systems rarely repeat the same decision class often enough to populate it. A baseline built per execution primitive or skill, in the manner already used in robot introspection research (switching autoregressive models over per-primitive signatures), is a more defensible starting point. Whichever construction is chosen, the

reference profile must be versioned and bound to the hardware, firmware, model and policy version in force, and anchored externally to the system it describes, so that the system under observation cannot silently redefine what counts as normal for itself. For end-to-end learned policies specifically, the concrete signals available to build and evaluate such a baseline (action-chunk consistency, direction-reversal rate) are discussed in Section 5.

The observation stage should, wherever the architecture allows it, be arbitration-anchored rather than latency-anchored. Signals available at points of arbitration, a safety filter's correction of a nominal command, a controller switch in a runtime assurance architecture, a re-planning event, a chattering or reversal pattern at the actuator level, tend to be available across a wider range of architectures, and to require less access to internal reasoning, than signals such as raw decision latency or an exposed confidence score, both of which were found to be weak or architecture-specific indicators on closer technical review.

5. Minimum Evidentiary Surface

A category C mechanism is only defensible if it does not become a second channel for extracting proprietary information. Two design commitments follow from this.

Architecture neutrality

No single signal is available across every architecture. Constraint activation, solver infeasibility and decision margin are naturally observable in optimisation-based control (model predictive control, control barrier functions) and in symbolic planners. Controller-switch events are naturally observable in runtime assurance architectures. Action-chunk consistency and direction-reversal rate are the closest analogues available for end-to-end learned policies, including vision-language-action models, where recent technical literature (2025-2026) has begun to show that some intuitively obvious signals, in particular raw velocity thresholds, are poor predictors of failure for continuous-action architectures, while behavioural instability measures at the action level perform better. A minimum evidentiary surface should therefore be defined as a small family of signal classes, with the applicable subset declared per architecture, rather than as one fixed variable list.

IP minimisation

The surface should not require chain-of-thought, complete reasoning traces, model weights, source code, hidden states, attention weights, or a complete proprietary policy. A tiered model of conformity, informally L0 (boundary-observable signals: actuators, essential sensors, execution events, timestamps, versions - for a model's own version, a

cryptographic hash of the weights, not the weights themselves, is one workable, non-prescriptive example of what "version" can mean without exposing intellectual property), L1 (arbitration-observable signals: safety interventions, controller switches, fallbacks, overrides), and L2 (signals a provider may voluntarily expose beyond L0 and L1: constraint pressure, decision margin, solver state), allows a provider to declare which level it supports without being forced into a single, maximal disclosure standard. The one variable found, on review, to sit closest to genuine intellectual property, and therefore the one most defensibly excluded from a mandatory core, is the internal pattern of access to a system's own policies or rule sets, which is architecture-specific and reveals structural information about the policy itself. The relevant EU legal framework for protecting that kind of information is Directive (EU) 2016/943 on trade secrets; nothing in this note asserts that the approach proposed here automatically satisfies that Directive's requirements, which would depend on implementation detail beyond the scope of this note.

Independence of the evidentiary layer

A provider may already maintain extensive internal logs, telemetry and diagnostic information, and those records can be highly valuable for engineering and safety purposes on their own terms. Evidence intended to support independent

post-event examination, however, calls for something different: preservation that is structurally separated from the sole control of the system or provider under examination. A record that only the examined party can access, retain or delete does not carry the same evidentiary weight as one held by an independent layer, however complete that record might otherwise be. This is a design requirement of the proposal set out in this note, not a documented legal one: none of the instruments reviewed in Section 7 requires this specific separation, and this note does not claim otherwise.

What this is designed to reconcile

The proposed approach is intended to preserve three interests simultaneously. Providers should remain free to develop, update and compete on proprietary models, architectures and policies without being required to expose internal reasoning, model weights or source code merely to support evidentiary examination, provided that the reference profile and evidentiary configuration applicable to a preserved event remain bound to the corresponding system version. Users, operators and affected parties should, at the same time, have a greater possibility that materially significant events are preserved in a form that can later be submitted to the authority or competent examiner entitled to assess them. The preservation mechanism itself makes no determination of

fault, liability or correctness. It preserves the conditions for judgment; it does not exercise the judgment.

What “conditions for judgment” means in practice

The conditions for judgment referred to above are not a promise that every preserved checkpoint answers every question. They are the bounded chain of elements that, when available under the Minimum Evidentiary Surface declared for a given system, can allow a competent authority or an authorised examiner to examine: the reference state against which a deviation was measured; the observed deviation itself; the model, software, policy and safety configuration in force at the time; the relevant sequence immediately before and after the event; whether a safety or control mechanism, override or fallback subsequently intervened; and whether the preserved record is consistent with, or diverges from, an account later provided by the system or provider. Which of these elements are actually available depends on the Minimum Evidentiary Surface declared for that system and use case; this note does not assert that every checkpoint necessarily supports all of them. Evidence availability is not itself a factual conclusion, an attribution of cause, or a determination of liability: it describes only what can still be independently examined after the event. The preserved material is therefore not a judgment produced by the evidentiary layer; it is material made available for subsequent

evaluation by whoever holds the lawful or otherwise competent authority to examine the event.

6. The Confident-Wrong Limitation

A category C mechanism built around behavioural deviation or elevated uncertainty has a structural blind spot that must be stated plainly, not minimised. A system that has converged on a systematically incorrect behaviour typically executes it with low variance and high apparent confidence, precisely because, from the system's own perspective, nothing about the situation is unusual. The absence of a trigger under such conditions does not establish the absence of error, risk, or harm.

This limitation is not a speculative concern specific to this hypothesis. It corresponds, at the level of an established and mature safety standard, to the distinction drawn by ISO 21448 (Safety of the Intended Functionality, SOTIF) between hazards arising from a system malfunction, where a safety mechanism can detect a fault, and hazards arising from a functional insufficiency, where the system behaves exactly as designed and specified, and there is, in the standard's own terms, no fault for a monitoring mechanism to detect.

A category C mechanism is accordingly not, and should not be presented as, an error detector, a universal incident detector,

a safety certification, or a correctness monitor. It is one of several possible mechanisms for selecting which events become candidates for evidentiary preservation, and it is bounded by the same limitation that any deviation-based or uncertainty-based safety mechanism carries.

7. Existing Regulatory and Technical Components

The following instruments were checked directly, against primary or near-primary sources, across two rounds of adversarial research. Each is classified by the type of trigger it uses, in the A/B/C taxonomy of Section 3.

European Union

- AI Act, Article 12: a technical capability requirement for logging, tied to risk identification, post-market monitoring and operational monitoring purposes. No deviation-based trigger; the obligation is that the system must technically allow recording, not that it must select which events matter.
- AI Act, Article 19: providers must keep automatically generated logs, to the extent under their control, for a period appropriate to the system's purpose and at least six months.
- AI Act, Article 73: reporting of serious incidents once a causal link, or a reasonable likelihood of one, is

established. Category A. The provider may not alter the system in a way that could affect a later evaluation of causes, before informing the competent authority.

- Machinery Regulation (EU) 2023/1230, Annex III §1.1.9 ("Protection against corruption") and §1.2.1(f) ("Safety and reliability of control systems"), independently verified by direct reading of the official EUR-Lex text (Official Journal of the European Union, L 165/47-48): §1.1.9 requires that hardware relevant to safety be protected against accidental or intentional corruption, that safety-critical software and data be identified and protected, that the software installed and necessary for safe operation be identifiable, and that evidence be collected of legitimate or illegitimate interventions on the software or of modifications to the software or configuration. §1.2.1(f) requires that the tracing log of data generated in connection with an intervention, and of the versions of safety software uploaded after the machinery is placed on the market or put into service, be enabled for five years after such upload, exclusively to demonstrate compliance with the Annex following a reasoned request from the competent national authority. "Tamper resistance" and "tamper evidence" remain this note's own paraphrase; the Regulation's own terms are protection against accidental or intentional

corruption and collection of evidence of a legitimate or illegitimate intervention. Together, §1.1.9 and §1.2.1(f) are the closest EU precedent to parts of the Evidentiary Metronome, though still intervention-anchored rather than deviation-anchored.

- Product Liability Directive (EU) 2024/2853, Article 9: court-ordered disclosure of evidence already in a defendant's possession, subject to necessity, proportionality and explicit protection for trade secrets (Art. 9(3)-(4)). Under Article 10(2)(a), a national court may presume the product defective where the defendant fails to comply with that disclosure obligation. This is a post-hoc, litigation-triggered mechanism. It regulates what happens to evidence that already exists once a claim is plausible; it does not require a system to capture anything specific in advance.
- Cyber Resilience Act, Annex I: logging of security-relevant internal activity. A different purpose (intrusion and vulnerability detection), not a behavioural or decisional mechanism.
- General Product Safety Regulation (EU) 2023/988: supply-chain traceability and outcome-triggered accident notification. Category A.

Technical standards

- ISO 21448 (SOTIF): voluntary, road-vehicle-specific, design and validation methodology. In the standard's own terms, it addresses hazards from functional insufficiency, distinct from the malfunctions addressed by ISO 26262. This note's framing of that distinction as "no fault for a monitoring mechanism to detect" is its own paraphrase of the malfunction/functional-insufficiency distinction, not a quotation from the standard, and should be read as such. It remains the clearest available independent confirmation, in a mature safety standard, of the confident-wrong limitation described in Section 6.
- ISO/IEC 42001 (AI management systems): independent compliance catalogues describe Annex A as including a control covering AI system operation and monitoring, including model drift, and a separate control for recording of event logs, though the catalogues consulted were not fully consistent on the specific clause numbering. Whether the standard specifies declared thresholds or automatic runtime alerts in the more specific sense used elsewhere in this note could not be confirmed against ISO's own primary text, which is paywalled, and should not be relied upon until checked. It remains, in any event, a voluntary organisational

management standard aimed at model governance and continuous improvement, not a technical product requirement.

- ISO/IEC FDIS 24970 ("AI system logging", ISO/IEC JTC 1/SC 42, project 88723): confirmed directly on iso.org to have reached Final Draft International Standard stage (stage 50.20, FDIS ballot initiated), a 26-page document, status "under development". The official ISO scope statement is limited to describing common capabilities, requirements and a supporting information model for logging of events in AI systems, designed to be used alongside a risk management system. This note found no official confirmation, from ISO, the European Commission, CEN or the Official Journal of the European Union, that ISO/IEC 24970 is already a published standard, that it is or will become a harmonised European standard, that it has been cited in the Official Journal, that it carries a presumption of conformity with AI Act Article 12, or that its content implements anything resembling Category C, a behavioural baseline, or deviation-triggered preservation. FDIS status is verified; Category C relevance is unverified. This remains the single most important open item for any future revision of this note.

United States, United Kingdom, China, South Korea

- United States: 49 CFR Part 563 (event data recorders, crash-pulse triggered) and NHTSA's Standing General Order for ADS and Level 2 ADAS crash reporting are both outcome-triggered (category A). No AI-specific federal equivalent to categories B or C was found.
- United Kingdom: no AI-specific binding logging or preservation requirement for AI systems in general. The UK's approach remains principles-based and sector-led, reconfirmed as of mid-2026 (no standalone AI bill following the May 2026 King's Speech). The Automated Vehicles Act 2024 (royal assent, sections not yet fully in force) is specific to automated vehicles rather than AI generally, and, confirmed directly on legislation.gov.uk and in the Bill's Explanatory Notes, provides for statutory inspectors who may investigate a "relevant incident" once it has occurred, with powers to require information, items and material, further provision on physical evidence (s.74), and protection of information (s.73), expressly not for the purpose of establishing blame or liability. This is a post-event investigative access and preservation mechanism, triggered by an incident that has already occurred, not a Category C evidentiary materiality trigger in the sense used in this note: nothing

in the Act evaluates a versioned behavioural reference profile against a runtime deviation.

- UNECE Regulation No. 157 and China's GB 44497-2024 (data storage system for automated driving, mandatory from 1 January 2026, following the same DSSAD structure) require recording not only crash-related data but a defined set of system-state events: activation and deactivation of the automated function, transition demands, driver overrides, and minimal risk or emergency manoeuvres. This spans both Category A (crash reconstruction) and Category B (discrete, already-defined system-state events) as used in this note, and this note's earlier characterisation of DSSAD as purely outcome-triggered has been corrected accordingly. It does not amount to Category C in the sense used here: the recorded events are enumerated in advance as fixed event types, not derived from evaluating a versioned behavioural reference profile against a runtime deviation. Separately, China's Ministry of Transport, in a national guide for automated driving transport services (Guide on Safety Services for Automated Driving Vehicle Transport (Trial), Doc. No. 交办运〔2023〕66号, issued 21 November 2023, released 5 December 2023), requires vehicles operating taxi, bus, passenger or freight services to record, on accident or automated-function failure, at

least 90 seconds of operating status data immediately preceding the event, including vehicle identification, control mode, location, motion state, environmental perception, light and signal status, 360-degree external video, and in-cabin video and audio of the driver and human-machine interaction. This is an event-triggered, pre-event buffer keyed to a crash or a system malfunction, structurally closer to Category A+B than to unconditional continuous recording; this note's earlier framing of it as closer to total recording has been corrected.

- South Korea: under the AI Basic Act and its implementing framework currently in force (the Act took effect from 22 January 2026, with its enforcement decree and the operative duties for high-impact AI reported to have taken full effect from 21 July 2026), operators of high-impact AI are subject to specified safety and reliability measures and to documentation requirements, including five-year retention of documentation evidencing those measures. This is documentation-based governance, structurally similar to the AI Act, not a deviation-triggered preservation mechanism, and is not classified here as Category C.

Industrial and academic precedent (not binding law)

- The Simplex Architecture and its black-box variant: a verified backup controller takes over from an unverified one when a formally verified safety condition is met. The clearest technical precedent for an arbitration-observable trigger, but the trigger is formally verified for a control-safety purpose, not statistically derived for an evidentiary one.
- Winfield and Jirotko's Ethical Black Box (2017; draft open standard, 2022): the closest named precedent in the robotics ethics literature. It is explicitly designed to record internal state data, including, on the authors' own account, the sequence of internal states of a system's ethical reasoning. This note reads that design choice as structurally opposite to the minimisation principle worked with here; that reading is this note's own comparison, not a claim made by Winfield and Jirotko.
- Tesla's disclosure-triggered data collection ("shadow mode"): independently reported, including by IEEE Spectrum, to run in parallel with human driving since 2016, comparing the system's predictions against the driver's actions and triggering the capture of a data snapshot (cameras, speed, acceleration and other parameters) when the two diverge. The closest functional industrial analogue to a deviation-selects-preservation

pipeline found anywhere in this research. Its baseline is human behaviour, not a self-referential system baseline, and its purpose is training data collection, not independent evidentiary preservation. It is reported to operate across Tesla's production fleet, though this note relies on industry reporting and public technical presentations rather than an audited specification.

CUSUM-based change-point detection, out-of-distribution detection in reinforcement learning, and recent (2025-2026) runtime failure-detection work on vision-language-action models (including conformal-prediction-calibrated methods) supply usable statistical building blocks for the observation and materiality-condition stages of the Metronome, but none of it is evidentiary law or standard in itself; it is cited here as available technical method, not as prior art for the regulatory function.

8. The Compositional Gap

Taken individually, none of the components of the Evidentiary Metronome is new. Version binding exists in the Machinery Regulation's tracing requirements for the safety-software versions covered by Annex III §1.2.1(f). A cybersecurity-driven tracing and retention obligation exists in the same instrument, and a narrower, post-incident duty not to alter a system exists in AI Act Article 73. Disclosure of existing evidence with trade-secret protection exists in the Product Liability Directive. Enumerated system-state event recording exists, for

automated driving specifically, in UNECE Regulation 157 and its national implementations. Operational monitoring covering model drift is described, in secondary sources, as part of ISO/IEC 42001. A technical logging capability exists in AI Act Article 12. None of this should be presented, or claimed, as new.

What was not found, in any binding instrument or standard checked across seven jurisdictions and blocks (the European Union, the United States, the United Kingdom, China, South Korea, together with UNECE automotive regulation), is these elements combined into a single function: a trigger declared in advance, applied independently of whether harm or fault is later established, based on a versioned behavioural reference rather than on an outcome or a prior safety intervention, that selectively preserves a bounded window around the event rather than requiring continuous total recording.

This is, on the evidence gathered so far, a compositional gap rather than a substantive one. The individual mechanisms exist. Their combination, for this specific evidentiary purpose, does not appear to. This distinction matters and should not be blurred in either direction: the gap is real enough to be worth a research question, and narrow enough that it should never be described as an absence of any relevant regulation or prior art.

9. Open Prior-Art Dependencies

The following items are not yet closed, and any conclusion in Section 8 should be treated as provisional until they are.

- The full technical content of ISO/IEC 24970 once published, and whether a CEN harmonised standard citing it is in fact adopted and cited in the Official Journal under AI Act Article 12. This is the single item most likely to change the classification in Section 8.
- The primary text of ISO/IEC 42001, Annex A, in particular the monitoring and event-logging controls that secondary compliance catalogues associate with drift and anomaly detection, currently known to this note only through those catalogues, which were not fully consistent with one another on clause numbering.
- IEC 61508 and ISO 26262, checked so far only at the level of general structure, not primary text, regarding whether either contains a diagnostic-coverage or safety-instrumented-function concept that amounts to a category C trigger.
- The IEEE standards portfolio beyond the social-robotics Ethical Black Box literature, and the NIST AI Risk Management Framework's traceability guidance, neither checked in depth.

- Canada and Japan, not yet checked with primary sources in this research.
- The Medical Device Regulation, relevant to medical robotics, not yet checked.

Until these are closed, Section 8's classification (a compositional, not substantive, gap) should be read as the best available conclusion given the sources checked, not as a final determination.

10. Research Question

The question this note leaves open, deliberately, is not whether to draft a technology-neutral requirement along the lines of a provider obligation to make available a minimum evidentiary surface capable of supporting the identification and preservation of materially significant decision-boundary events, without requiring disclosure of internal reasoning, model weights, source code, or other information not necessary for later independent examination. No such text is proposed here.

The question is narrower and prior to that: whether the compositional gap identified in Section 8 is significant enough, and stable enough against the dependencies listed in Section 9, to justify moving from research to a defined next step, and if so, which step. Candidates, in ascending order of commitment, include: further narrowing of the hypothesis to a single architecture class for a first technical validation; a

falsification experiment along the lines already scoped in earlier technical review, comparing a materiality signature's ability to separate genuine competing-constraint scenarios from confounder scenarios; closer tracking of ISO/IEC FDIS 24970 pending its publication; or the drafting of a technology-neutral discussion text for circulation to a small number of external reviewers, stopping short of a legislative proposal.

This note does not resolve that choice. It records, as precisely as the available sources allow, what is already covered, what is adjacent but different in kind or purpose, and what, on the evidence gathered so far, still appears to be open.

Appendix: Element-by-Element Summary (A-K)

Elements A to K correspond to the components of a hypothetical category C mechanism, tested individually against the sources reviewed in Section 7. Status labels: Required, Partially required, Technically known (documented in the technical or academic literature but not standardised or mandated), Not found.

Element	Status	Best source found
A. Minimum evidentiary surface	Partially required	AI Act Art. 12(1)-(2) (generic capability); ISO/IEC FDIS 24970 (FDIS stage verified on iso.org, content unverified, link to Art. 12 presumption of conformity unconfirmed)
B. Reference profile / baseline	Not found	No instrument verified requires a versioned behavioural baseline
C. Deviation / arbitration observation	Partially required (Category A+B, not C)	UNECE Reg. 157 / GB 44497-2024 DSSAD - enumerated system-state events (activation, transition demand, override, minimal risk manoeuvre); Machinery Regulation (EU) 2023/1230, Annex III §1.1.9 (corruption protection) and §1.2.1(f) (tracing log on intervention), verified on EUR-Lex; ISO/IEC 42001 Annex A (secondary catalogues; primary text unverified). None constitutes

		a versioned baseline evaluated against runtime deviation
D. Blame-neutral materiality condition	Not found	All verified triggers are outcome-based (AI Act Art. 73, GPSR) or enumerated system-state events (UNECE R157)
E. Automatic trigger from deviation	Not found	No verified instrument
F. Selective pre/post-event preservation	Not found (for Category C)	Closest functional analogues are outcome/malfunction-triggered pre-event buffers, not deviation-triggered: China MOT Guide (Trial), Doc. 交办运〔2023〕66号 (>=90s pre-event buffer on accident or automated-function failure); Tesla shadow mode (industrial, not regulatory)
G. Minimum evidentiary set	Partially required	Machinery Regulation (EU) 2023/1230, Annex III §1.2.1(f), independently verified (OJ L 165/48): tracing log of intervention data + safety-software versions, enabled for five years; UNECE R157/DSSAD and China MOT Guide (enumerated event sets); AI Act Art. 12(3) only for real-time biometric identification
H. Integrity / tamper resistance	Partially required	Machinery Regulation (EU) 2023/1230, Annex III §1.1.9, independently verified (OJ L

		165/47-48): protection against accidental or intentional corruption, collection of evidence of a legitimate or illegitimate intervention ("tamper resistance/evidence" is this note's paraphrase of those terms, not the Regulation's own wording); AI Act Art. 73(6) (no alteration before notifying authorities)
I. Version binding	Required	Machinery Regulation (EU) 2023/1230, Annex III §1.2.1(f), independently verified (OJ L 165/48): tracing of safety-software versions uploaded after placing on the market or putting into service, enabled for five years, for the purpose of demonstrating compliance on reasoned request; scope limited to machinery within the Regulation, not generalised to AI systems at large
J. Independent post-event examination	Partially required	Market surveillance access to logs; PLD Art. 9 court-ordered disclosure; UK Automated Vehicles Act 2024, ss.60-74 (statutory inspectors, post-incident, blame-neutral by statute, not Category C)
K. IP / trade-secret minimisation	Partially required	PLD (EU) 2024/2853, Art. 9(3)-(4)

	(litigation stage only)	
--	-------------------------	--

This table reflects the sources checked as of the date of this note. It should be re-run against ISO/IEC FDIS 24970 once published, and against the jurisdictions and standards listed in Section 9 as still open, before being relied on as final.

Author Note

Dr. Emanuel Celano

Founder & Chief Architect, EVIDE

This Research Discussion Note was developed as part of EVIDE's ongoing research into independent evidentiary architectures for AI and autonomous systems.

The concepts proposed in this paper - including the **Evidentiary Metronome**, **Evidentiary Materiality Condition**, **Evidentiary Trigger**, **Evidentiary Checkpoint**, and **Minimum Evidentiary Surface** - are presented as research hypotheses for technical, legal, regulatory and industry discussion.

The objective is not to prescribe how AI systems should make decisions, but to explore how evidence can be preserved at materially significant moments **without requiring unnecessary exposure of proprietary systems and without assigning error, blame or responsibility at the point of preservation.**

Visit EVIDE:

<https://app.certifywebcontent.com>