From Record to Finding

Why Tamper-Proof Logs Cannot Establish Legal Oversight of Agentic AI

Jeroen Janssen
Apparens
Deventer, Netherlands · jeroen@apparens.nl

Working Paper
June 2026

This is a working paper circulated for discussion and comment. It has not been peer reviewed. Comments are welcome.

DOI: 10.5281/zenodo.21025237

Abstract

A supervisor charged with overseeing an agentic AI system must, under EU law, be able to establish certain findings of fact: whether personal data of a protected class left a controlled environment, whether a human had the authority and the window to intervene, whether an information barrier held, whether a delegated authority was valid at the moment it was used. This paper asks a narrow and prior question. Under what conditions can a runtime record *answer* such a finding at all? It develops an evidentiary-adequacy criterion: a representation can answer a determination of this kind only if it carries two things, the legal typing that maps recorded events to the operative category, and the relation, usually provenance or authority, on which the determination's truth depends. A representation carrying neither can raise a suspicion but cannot establish a finding. The criterion is a necessity claim, not a sufficiency claim; sufficiency fails for reasons the paper makes explicit. The criterion is stated formally, its necessity is shown by construction across a defined class of determinations, the property set is shown to be minimal, and its adequacy is argued to be bounded to that class rather than universal. The determinations in scope are characterised by a stated selection rule: they are binary findings of fact about specific events and their relations, not evaluative or statistical properties such as fairness or robustness. The criterion is supported, not derived, by three independently established results that converge on the same necessity from different directions: Ashby's Law of Requisite Variety, the Good Regulator Theorem, and the trace-versus-hyperproperty boundary of runtime verification. The paper is careful to claim no novelty in that convergence; the inversion of cybernetic model-dependence onto the overseer has been made before, notably by Aguirre (2025), as an impossibility argument about control. The contribution is the criterion itself, its instantiation in EU AI Act oversight, a gap analysis showing that existing governance frameworks do not answer these determinations, and a pre-registered experiment that tests the criterion against expert judgement. The legal claim is stated conditionally. A record that is not semantically and relationally interpretable cannot serve as evidence that effective oversight under Article 14 was possible, even where it satisfies the Article 12 duty to log. The structure recurs in any domain where legal findings depend on the execution paths of autonomous systems, of which EU AI Act oversight is the worked instance.

1. Introduction

An agentic AI system is software that uses a model to decide what to do and a harness to do it. One property distinguishes it from ordinary software. Its execution path, the sequence of steps it takes, is selected at run time and cannot be enumerated in advance, and that selection can be steered by the content the agent reads. A conventional service fixes its control flow at design time. An agent composes its control flow as it runs.

The dominant response in enterprise governance has been to extend the instruments built for conventional software: the point-in-time audit, the conformity checklist, the periodic review. These are static. They assess a system at a moment, against fixed criteria, and produce a record that is fixed once produced. A familiar line of argument, which this paper develops but does not rest its contribution on, holds that a static instrument is in the wrong structural class to oversee a system whose behaviour is generated rather than fixed.

This paper makes a different and prior move. Before asking whether oversight succeeds, it asks what a record must contain for a supervisor to read a legally operative fact from it at all. That is an evidentiary question, not a control-theoretic one, and it has a precise answer for a well-defined class of facts.

The contribution, stated up front. The paper defines an *evidentiary-adequacy criterion* for runtime oversight. For a class of legal determinations that are findings of fact about specific events and their relations, a runtime record can answer the determination only if the record carries two features: a *typing* that maps recorded events to the legally operative category, and a *relation*, typically provenance or authority, on which the determination's truth depends. The criterion is a necessity claim. It is shown by construction for each determination in the class, the two-feature property set is shown to be minimal, and the criterion's adequacy is argued to be bounded to the defined class. Sufficiency is explicitly out of scope and is shown to fail. The criterion is then instantiated against the oversight obligations of the EU AI Act, supported by three convergent results from cybernetics and computer science that the paper is careful not to claim as novel, and tested by a pre-registered experiment.

What this paper does not claim. It does not claim that the cybernetic inversion, applying model-dependence to the overseer rather than the AI, is new. That move has been made, and made well, by Aguirre (2025) as an impossibility argument about the control of superintelligence. It does not claim that a particular ontology is required. It does not claim that the criterion is sufficient for good oversight. It does not claim that requisite variety and runtime monitorability are the same theorem. Each of these restraints is load-bearing and is observed throughout.

Research question. Under what conditions can a runtime record answer a legally operative determination about an agentic AI system, and what minimal structure must the record carry for that to be possible?

Why it matters beyond AI governance. The structure is not specific to AI. Wherever a legal finding about an autonomous system depends on the system's execution path, on the relation between actions, or on the provenance of a flow, the same criterion applies: autonomous finance, medical devices, critical infrastructure, autonomous defence, distributed software agents. EU AI Act oversight is the worked instance in this paper because its obligations are explicit and current. The criterion is stated so that the instance can be lifted.

Roadmap. Section 2 defines the criterion formally, states the selection rule for the determinations in scope, proves necessity by construction and minimality, and bounds completeness. Section 3 shows by gap analysis that existing governance frameworks do not answer these determinations. Section 4 presents the three convergent supporting results and names the prior work that reaches each. Section 5 frames oversight as a control loop and locates the sensing failure. Section 6 works the information-barrier example. Section 7 instantiates the criterion in EU AI Act oversight, in conditional terms. Section 8 separates the descriptive and normative models and states falsifiable predictions. Section 9 gives the pre-registered experiment. Section 10 is the critical evaluation. Section 11 draws implications and the cross-domain generalisation.

Appendix A is the claim inventory and falsification register. Appendix B is the experiment protocol.

A note on terms. Four words are used in fixed senses throughout, and the paper does not let them drift. A *model*, except where a cited result fixes a narrower technical sense, means a representation adequate to recover a determination's truth value. A *regulator* means a function coupled to a system in a feedback relation, sensing, deciding, and acting on the system's future states; a function that only inspects records is an *evaluator*, not a regulator. *Sensing* means the recovery of a determination's truth value from a record, the evidentiary counterpart of perception. *Evidence* means a record from which a legally operative fact can be recovered, as distinct from a record that merely exists and has integrity. Where a cited theorem uses one of these words in its own technical sense, the paper flags the difference rather than trading on it.

2. The Evidentiary-Adequacy Criterion

This section states the contribution before any borrowed result is invoked. The criterion stands on the argument given here. Sections 4 and 5 corroborate it from cybernetics and computer science, but the criterion does not depend on them, and nothing in this section is inferred from a cybernetic theorem.

2.1 Determinations, evidence, answerability

Let a *determination* D be a legally operative finding of fact that a supervisory system must be able to establish about an agentic system's behaviour. Let an *evidence representation* E be the record, in whatever form, from which the supervisory system attempts to establish D. Let an *answering procedure* be a decision procedure f that takes E and returns a truth value for D.

Definition (answerability). E *answers* D when there exists a decision procedure f such that f(E) recovers the truth value of D, soundly, for the cases in the determination's scope. E *fails to answer* D when no such procedure exists over E, that is, when the truth value of D is not a function of the content of E.

This is a deliberately modest definition. It says nothing about how E is produced, whether the supervisor is human or automated, or whether oversight as a whole succeeds. It isolates one question: is the fact recoverable from the record, or is it not present in the record to be recovered?

2.2 Which determinations: the selection rule

The criterion does not range over every property one might wish to establish about an AI system. It ranges over a specific class, and the class is defined by a stated rule rather than by example, so that the selection cannot be read as cherry-picking.

Selection rule. A determination is *in scope* when it is a binary finding of fact about specific events and their relations: a yes-or-no question whose truth value is fixed by what specific events occurred, what authority they ran under, and how they relate to one another. A determination is *out of scope* when it is an evaluative or statistical property over a distribution of behaviours rather than a finding about particular events.

The four determinations used throughout are in scope by this rule: whether protected data crossed a boundary, whether a human held authority and a usable window to intervene, whether an information barrier held, whether a delegated authority was valid when used. Each is a yes-or-no fact about specific events and their relations.

Fairness, robustness, non-discrimination, explainability, and transparency are out of scope. They are properties over distributions of outcomes, not findings about a particular event, and they raise representational questions the criterion does not address. Excluding them is not a weakness of the criterion; it is the boundary that makes the criterion precise. A reviewer who asks "why not fairness" is answered by the rule: fairness is not a finding of fact about an event, so it is a different kind of object requiring a different treatment.

2.3 The criterion

Criterion (evidentiary adequacy). Let D be a determination in scope under 2.2, of the form "did event-type X, of legal category C, stand in relation R to event-type Y." Then E answers D only if E represents, explicitly or in functionally equivalent form, both:

1. a *typing* that maps the recorded events to the legal category C, and

2. the *relation* R, usually provenance, derivation, or authority, on which the truth of D depends.

A representation E that carries neither the typing nor the relation can, at most, detect that something occurred or that something looks anomalous. It cannot establish D as a finding of fact.

The criterion is stated as a necessary condition. It is not claimed to be sufficient. Section 2.8 states why sufficiency fails and places it out of scope.

2.4 Findings, not anomalies

The line the criterion draws is sharper than the difference between a good record and a bad one, and it must be stated exactly because the commercial temptation is to blur it. A behavioural anomaly detector can flag that something looks wrong without representing the protected category or the relation. It raises a suspicion, not a finding. The determinations EU law places on a supervisor are findings: that protected data did or did not cross a boundary, that a barrier did or did not hold. Establishing a finding, rather than raising a suspicion, is precisely what requires the typing and the relation. The criterion is a criterion for findings. It says nothing about the detection of risk, which can proceed on weaker evidence and serves a different purpose.

2.5 Necessity by construction

The criterion's necessity is not asserted; it is shown, determination by determination, by exhibiting a bare action log and demonstrating that the truth value of the determination is not a function of the log's content. A bare action log is taken to record, for each action, that it occurred, when, and that it was permitted, with no typing of the data or channel and no representation of derivation or authority state.

Table 1. Necessity by construction: why a bare log cannot answer each determination.

Determination D	Truth value fixed by	What the bare log contains	Why D is not recoverable from the bare log	Minimal addition that makes D recoverable
Did protected data leave the controlled environment?	Whether data of a protected class was the content that crossed an external boundary	That a send occurred and was permitted	The log does not represent the class of the data sent, nor whether the channel was internal or external; both are needed and neither is present	Typing of data by protected class; typing of the channel as internal or external
Was a human able to intervene?	Whether a human held authority and a usable window to override at the decision point	That an action executed	The log does not represent whether an override path existed or was reachable in the available time	Representation of the authority state and the intervention window at the decision point
Did the information barrier hold?	Whether information derived from a high-side source reached a low-side sink	That two permitted actions occurred	The log records two actions but not the derivation relation between them; the truth of D is the relation, which is absent	Provenance linking the send to the source read; typing of source and sink by barrier side
Was a delegated authority valid when used?	Whether the authority was granted, unrevoked, and in scope at the instant of use	That the action occurred	The log does not represent the validity state of the authority at that instant	Temporal model of delegation and revocation; binding of the action to the authority state

In each row the truth value of D depends on a typing and a relation that the bare log does not carry. The determination is therefore not a function of the log's content, and by the definition in 2.1 the log fails to answer it. This is a constructive negative result, not an appeal to intuition.

2.6 Minimality

The two-feature property set is minimal: neither feature can be dropped without losing the ability to answer some determination in the class.

Drop the typing. Then for the first determination, the log may record that a send crossed an external boundary, with provenance, but cannot say that the content was of a protected class. The finding "protected data left the environment" is not recoverable. Typing is necessary.

Drop the relation. Then for the third determination, the log may type each action, recording that a deal-room read occurred and that a desk send occurred, but cannot represent that the send was derived from the read. The finding "the barrier held" is not recoverable, because the barrier is the relation. The relation is necessary.

Neither feature is redundant. The set {typing, relation} is necessary and minimal for the class.

2.7 Bounded completeness

A natural question is whether the two features are *enough* to answer every determination in the class, or whether some in-scope determinations need more. The honest answer is that completeness is claimed only for the class as defined, not universally, and that the relation primitive must be read broadly enough to carry what the determinations require.

Two of the four determinations turn on time and on authority state, not on derivation alone. "Was a delegated authority valid when used" requires a temporal model of grant and revocation; "was a human able to intervene" requires the authority state and the window at the decision point. These are carried by the *relation* feature only if "relation" is read to include authority and temporal relations, not provenance alone. The paper reads it so, explicitly: the relation feature ranges over provenance, derivation, authority, and the temporal ordering on which validity depends. Under that reading the two-feature set is adequate for the four determinations and for determinations of the same form.

The bounded claim is therefore: {typing, relation-broadly-construed} is necessary and minimal for the class in 2.2, and adequate for it, where "relation" carries provenance, authority, and temporal structure. The paper does not claim adequacy for determinations outside the class, and it does not claim that no further feature could ever be needed for some in-scope determination not yet considered. Completeness over everything is the overreach the paper declines, for the same reason it declines an if-and-only-if theorem in Section 10.

2.8 Sufficiency is out of scope, and fails

The criterion is necessity, not sufficiency. A representation that carries the typing and the relation can still fail to support good oversight, and it can fail in a way the criterion cannot detect. Section 10.6 develops this through Power's (2007) account of performative compliance. A semantic representation buys legibility, not truthfulness: it makes a record readable for a determination; it does not make the record honest. A legible record can be a curated one, typed to show compliance and to omit what would show its absence. Sufficiency therefore depends on facts about the deployment, the good faith of the supervised, and the competence of the supervisor that no representation can supply. The paper claims necessity and disclaims sufficiency, and the disclaimer is not a hedge but a boundary that a sufficiency claim would cross.

2.9 Equivalent representations

The criterion is satisfied by structure, not by a particular schema. The typing and the relation may be carried as labels, classifications, taint tags, lineage metadata, correlation rules, proof circuits, policy graphs, or an ontology. The paper does not claim that any one of these is required, and it does not claim uniqueness of any representation. It claims only that *some* representation of the typing and the relation must be present, in explicit or functionally equivalent form. A data-lineage model that tracks derivation is carrying the relation. A procedural model that records which events count as a barrier

crossing is carrying the typing. The disagreement with an objector who proposes any such mechanism is over the name, not the substance: whatever the mechanism is called, to answer "did the barrier hold" it must represent the protected class and the derivation relation, and a mechanism representing neither cannot answer it. This is why the contribution is a criterion over representations and not a proposal for one.

3. Why Existing Governance Frameworks Do Not Answer These Determinations

A reviewer is entitled to suspect that the criterion is already satisfied by some framework in wide use, and that the contribution is therefore a relabelling. It is not. The frameworks that govern AI systems either produce records without the required structure, or specify processes rather than evidence. The distinction the criterion draws, between producing a log and producing a finding, is exactly the distinction these frameworks do not make.

Table 2. Existing frameworks against the answerability criterion.

Framework	Produces records or logs?	Produces findings of fact about events and relations?	Why not, against the criterion
NIST AI Risk Management Framework	No direct record duty; a process and outcomes framework	No	It governs risk-management process and organisational function; it does not specify an evidentiary representation that types events and carries relations
ISO/IEC 42001 (AI management systems)	Through the management system, indirectly	No	It certifies that a management system exists and operates; it does not specify the content of a runtime record from which a legal finding is recoverable
EU AI Act, Article 12 (logging)	Yes, automatic event recording over the lifetime	Not on its face	The recording duty requires events be logged for traceability; it does not require that the log type events to legal categories or carry the relations on which findings depend (see Section 7)
Runtime verification and enforcement	Yes, as a by-product of mediation	Partially, and only with instrumentation	It mediates and may halt actions, and can produce enriched traces; but a black-box single-path monitor cannot evidence relational properties without the system knowledge the criterion calls a model

Framework	Produces records or logs?	Produces findings of fact about events and relations?	Why not, against the criterion
			(see Section 4.3)
Provenance models (PROV and successors)	Yes, derivation graphs	The relation, but not the legal typing	Provenance carries the relation feature; on its own it does not type events to the legal category C, so it answers "was Y derived from X" but not "was X of protected class C"

The pattern across the table is the point. Process frameworks (NIST, ISO/IEC 42001) do not produce findings because they govern function, not evidence. Record frameworks (Article 12, generic logging) produce records without the typing and relation, so they produce data and not findings. The components that carry one feature (provenance carries the relation; typing schemes carry the category) do not on their own carry both. No framework in current use produces, by itself, a record that satisfies the criterion for the determinations in scope. That gap is what the criterion names and what a supervisory-evidence representation must close.

4. Three Convergent Supporting Results

The criterion in Section 2 is argued from the nature of the determinations and does not depend on any borrowed theorem. This section adds corroboration. Three results, established independently in cybernetics and computer science, converge on the same necessity from different directions. The paper claims no novelty in the convergence and names, for each result, the prior work that has already applied it to oversight. The results are presented as lemmas that support the criterion, not as its source.

4.1 Requisite variety: static audit as the sole mechanism is in the wrong class

Ashby's Law of Requisite Variety (Ashby, 1956) states that a regulator can hold an essential variable in range across a set of disturbances only if it can produce at least as many distinct compensating responses as there are distinct disturbances to compensate. A distinction internal to the law matters here. A regulator's *intrinsic* variety is the set of responses it is in principle capable of; its *effective* variety is the subset it will actually deploy given its decision rule and constraints, and the law is stated over effective variety. Aguirre (2025, footnote 100) makes the same distinction in the oversight setting, observing that requisite variety applies not to the controller's internal complexity but to the effective variety it can express through its control signals per unit time.

Applied to oversight, the law yields a structural reading, qualitative and not a derivation from information theory. A point-in-time audit, *taken as the sole oversight mechanism*, is a regulator that acts once: it senses at a moment, applies fixed criteria, and produces a record fixed once produced. Its effective variety is exhausted on acting. A path-selecting agent fleet generates variety continuously. A regulator that acts once, against a system that acts continuously, is in the wrong variety class to hold a runtime-dependent

essential variable in range, and adding detail to the checklist raises its intrinsic variety without raising the effective variety it can deploy after its single action.

The qualifier "as the sole mechanism" is essential and is stated, not smuggled. The claim is not that audit has no place; a static audit can establish that a control environment exists, which is a real and necessary function. The claim is narrower: audit alone is in the wrong class to be the mechanism that holds a runtime-dependent variable in range. A defender who answers that runtime controls, monitoring, and assurance together form the regulator concedes the point, because that composite is no longer a static audit but a closed loop with a runtime sensing channel, which is what the paper argues the task requires.

To make this an application rather than an analogy, the regulator mapping must be specified against Ashby's conditions and not merely asserted. A regulator in Ashby's sense has an essential variable held in a range, a disturbance set, a sensing channel, a decision rule, an intervention repertoire, and a feedback latency. A legal condition such as "the information barrier holds" is an essential variable only when it is rendered measurable, sensed through a channel, connected to a decision rule, and tied to an intervention that can hold it in range. A legal condition that is none of these is not an essential variable; it is a legal condition relabelled in cybernetic vocabulary, and the objection that the apparatus is borrowed would, in that case, be correct. The work of an oversight design is to make the legal condition measurable, sensed, decisionable, and connected to intervention. Where that work is not done there is no regulator and no essential variable, only a record. The variety reading therefore corroborates the criterion exactly where the criterion already bites: a record that cannot be read for the operative fact cannot be the sensing channel of any regulator.

The relation of this variety reading to the monitorability limit of 4.3 is treated as a convergence, not an identity. Ashby framed requisite variety as a restatement of Shannon's channel-capacity results, and the information-theoretic limits of feedback control have since been made quantitative (Touchette and Lloyd, 2000). Aguirre (2025, Appendix A.5) builds a channel-capacity formalisation of the overseer's information rate on exactly this basis, comparing the rate at which a system generates choice-complexity to the rate at which an overseer can transmit constraint. That formalisation is the route by which the variety reading could be made quantitative; this paper does not claim to have carried it through, and notes that Aguirre has already developed it for the control setting. Whether the variety deficit of static oversight and the single-trace monitorability limit are the same constraint, rather than two pointing the same way, remains an open question, marked as conjecture in Section 10.3.

4.2 The Good Regulator Theorem: model-dependence in the coupled loop

Conant and Ashby (1970) proved that any regulator that is maximally both successful and simple must be modelled on the system it regulates; the gloss usually given is that every good regulator must be a model of its system. The theorem is frequently asked to carry more than it proves, and this paper takes it only for what it proves, under a stated narrowing.

First, the theorem's reach. Conant and Ashby assume a regulator coupled to its system in a feedback relation, one whose actions affect the system's future states. The theorem speaks to regulators in that closed-loop relation. It does not speak to a function that only inspects records after the fact and cannot alter future behaviour. Such a function is an evaluator, not a regulator, and the theorem's model requirement does not bind it. This is decisive for the paper's object and is stated as a limit rather than worked around: the cybernetic corroboration reaches the supervisory *system* that combines sensing, decision, and an intervention repertoire coupled to the path, and it does not reach a purely retrospective audit standing outside the loop. Where oversight is purely retrospective, the variety and theorem readings do not apply, and the criterion of Section 2, which is an evidentiary claim and not a control claim, carries the case alone. A search of the formal literature finds no credible extension of the theorem to an uncoupled retrospective evaluator.

Second, the theorem's strength. The formal result establishes a morphism between system states and regulator states, a model-dependence condition: a coupled regulator must carry a mapping adequate to select its action. It has been argued that the notion of model this supports is permissive, applying even where a system uses no model in a meaningful representational sense (Thobani, 2024; Virgo et al., 2025). Virgo et al. go further and relocate the model to the observer, treating it as imposed from outside rather than contained within. The paper does not rely on the strong representational reading and does not need it. It takes the theorem only for model-dependence: a coupled regulator must carry some mapping adequate to choose its action, and a regulator carrying no mapping is reduced to reacting to surface signals it cannot interpret. The further content, that for legal determinations the mapping must be a typed, relational representation, is not drawn from the theorem. It is the criterion of Section 2, argued from the determinations. The theorem locates a dependence; the criterion specifies its content for the legal case; the two are kept separate so that the borrowed result is never asked to carry the legal conclusion.

This inversion, applying model-dependence to the overseer rather than to the AI, is not novel, and the paper is explicit about the nearest prior work. Aguirre (2025) applies both requisite variety and the Good Regulator Theorem to the overseer of a superintelligence, observes in his own terms that the theorem is weaker than its slogan and must be modified to yield a model requirement, and builds the modeling obstacle and a channel-capacity formalisation of the overseer's limits. His purpose is the opposite of this paper's: he runs the inversion as an impossibility argument at civilizational scale, within the loss-of-control frame. Notably, Aguirre himself brackets oversight, "the monitoring and after-the-fact correction of AI behavior," as a concept distinct from and weaker than control, and sets it aside (Aguirre, 2025, footnote 3). This paper takes up exactly what he sets aside, on the oversight side of his own line, and at enterprise scale within the EU-law evidentiary frame. The cybernetic move is shared and disclaimed as prior; what this paper adds is on the evidentiary and legal side, not the cybernetic side.

A second strand of prior work treats audit itself as a regulatory function, and must be named because it is the nearest precedent on the audit side. Beer's Viable System Model casts its audit channel, System 3*, as part of organisational regulation rather than as a detached inspection, and cybernetic audit theory develops the point that auditing is a

trust-creation process within a regulated system (Beer, 1979; Espejo, 2001). That tradition already reads audit cybernetically. It does not connect audit to agentic execution traces, to the trace-versus-hyperproperty boundary, or to the evidentiary requirements of EU law, and it does not state an answerability criterion over determinations. The contribution of this paper sits in the conjunction those two strands leave open: the cybernetic-audit tradition reads audit as regulation but not over agentic evidence, and the overseer-inversion tradition reads oversight as model-dependent but not as a legal-evidentiary criterion. Neither reaches the criterion of Section 2, and the criterion does not depend on either.

4.3 Runtime monitorability: from impossibility to an instrumentation requirement

The computer-security lineage supplies a third convergence, and the honest reading of it supports the criterion rather than a blunt impossibility.

Schneider (2000) defined the class of security policies enforceable by a monitor that observes a single execution and may halt it, and showed this class is contained in the safety properties. Clarkson and Schneider (2010) distinguished trace properties, which are sets of executions, from hyperproperties, which are sets of sets of executions, and showed that noninterference, the property that high-confidentiality inputs have no observable effect on low-confidentiality outputs, is a 2-safety hyperproperty: a violation is witnessed only by a pair of executions, never by one alone.

It is tempting to state this as "a single-path monitor cannot evidence a hyperproperty," full stop. That statement is true only in a trivial and self-defeating sense, and the paper does not rely on it. A 2-safety hyperproperty is refuted by exhibiting two executions; it reduces to a safety property of the self-composition of the system. The correct reading is not impossibility but an instrumentation requirement: relational evidencing requires either more than one run, or a product construction, or knowledge of the system beyond a single black-box trace. The hyperproperty-monitoring literature establishes exactly this. Stucki, Sánchez, Schneider and Bonakdarpour (2019) show that black-box monitoring of HyperLTL is in general unfeasible, and propose a gray-box approach that performs static analysis of the system at run time, applying it to a privacy hyperproperty, distributed data minimality. This cuts in the paper's favour. The condition under which the relational property becomes checkable is precisely that the monitor be given knowledge of the system, which is the model-dependence condition restated in the security vocabulary, and which is the criterion's typing and relation made operational. The honest statement is therefore: a single-path black-box monitor cannot evidence a relational property, and the knowledge that lifts the limit is exactly a model of the system, carrying the typing and the relation.

4.4 Convergence, not identity

The three results point at the same gap from three directions. Requisite variety says a once-acting regulator cannot match a continuously-acting generator of variety. The Good Regulator Theorem says a coupled regulator must carry a mapping of its system. Runtime verification says relational evidencing requires knowledge of the system beyond a single trace. Each, applied to oversight, says that a bare record is not enough and that what closes the gap is a representation of the system's behaviour. None of the

three is the criterion, and the paper does not claim they are one theorem. They corroborate a criterion that is argued independently from the determinations. That separation is what keeps the contribution honest: if all three borrowed results were withdrawn, the criterion of Section 2 would stand on its construction and minimality proofs, and only the corroboration would be lost.

5. The Oversight Loop and the Sensing Failure

Oversight, stripped to its function, is a control loop. A supervisory system senses the behaviour of the system it oversees, decides whether that behaviour is acceptable, and intervenes when it is not. The word "system" is load-bearing. Under the EU AI Act, oversight is not discharged by a single person or artifact. It is distributed across provider design duties, deployer monitoring, the human operators who exercise Article 14 oversight, the logging infrastructure, post-market monitoring, and organisational governance. The claim of this paper is about the supervisory system as a whole: the configuration of records, schemas, tools, procedures, trained humans, and intervention rights must, taken together, contain or have access to a representation adequate to answer the determinations. Where the paper writes of what "the supervisor" must contain, it is shorthand for that distributed system.

Two mechanisms operate on the execution path. The first acts before an action runs, permitting or blocking it: an intervention mechanism, of which runtime enforcement systems are instances (Wang, Poskitt and Sun, 2025; Wang et al., 2025), descended from the reference monitor (Anderson, 1972; Saltzer and Schroeder, 1975). The second produces the record by which the supervisory system later determines what the agent did: a sensing mechanism, the system's instrument of perception. The intervention half has advanced. The sensing half has not advanced at the same rate, and the criterion of Section 2 explains why the gap matters: an intervention mechanism produces, as a by-product, a record, and where that record is hash-chained it is tamper-evident. Tamper-evidence establishes integrity, that the record was not altered after the fact; the eIDAS framework gives a qualified electronic timestamp a presumption of integrity of the bound data (Article 41(2)). Integrity is necessary for a record to serve as evidence. It is not sufficient. A record can have integrity and still fail to answer a determination, because integrity is a property of the record's history and answerability is a property of the record's content.

This yields a taxonomy of oversight failure by loop function. The sensing failure is the one the criterion isolates, and it is the one most easily concealed, because a record exists and appears complete.

Table 3. The oversight loop and its failure modes.

Loop function	Role in oversight	Failure mode	Manifestation for agentic AI
Sensing	The supervisor perceives what the system did	The record is present but cannot answer the determination at hand	A tamper-evident action log that records that an action occurred and was permitted, but not the legally operative fact

Loop function	Role in oversight	Failure mode	Manifestation for agentic AI
			the supervisor must establish
Decision	The supervisor judges behaviour against a standard	The standard cannot be applied because the relevant property is not a property of any single observed execution	An information-barrier breach visible only in the relation between two agents' paths, not in either path alone
Intervention	The supervisor stops or corrects behaviour	The intervention arrives after the behaviour has produced its effect, or cannot be targeted because the behaviour was not sensed	A periodic review that detects a pattern months after the actions that constituted it, when correction is no longer available

The literature on human oversight has concentrated on decision and intervention: on whether a human has genuine authority to override, on when oversight collapses into rubber-stamping, and on the assignment of responsibility (Chiodo et al., 2026, whose framing is consistent with the three-function loop though developed through computational reductions rather than a control loop as such). The sensing function has received less attention as a distinct failure mode. Yet sensing is prior. A supervisor who cannot perceive what the system did cannot judge it and cannot target an intervention. The first failure of oversight is a failure to sense, and the criterion is the condition under which that failure does not occur.

6. The Information-Barrier Example

The criterion can be made concrete with the smallest case that exhibits it.

Two agents serve one organisation. Agent A has read access to a deal room containing material non-public information. Agent B has authority to send messages to a trading desk. The organisation operates an information barrier: information from the deal room must not reach the desk. Each agent's path is individually compliant. Agent A reads the deal room, which it is permitted to do. Agent B sends to the desk, which it is permitted to do. No single action violates a per-action policy. The violation, if there is one, is the relation: information derived from A's read reaching B's send.

How that relation is classified admits more than one formulation, and the distinction does not rescue the bare trace; it reinforces the criterion. In its strongest confidentiality form the barrier is a noninterference property between a high-confidentiality source and a low-confidentiality sink, which by Clarkson and Schneider (2010) is a 2-safety hyperproperty, witnessed only by the relation between two executions, and which a single-path black-box monitor cannot evidence. In a weaker and operationally common form the barrier is an explicit-flow or taint property: data derived from the deal room must not reach the desk channel. A taint property can be monitored on a single enriched execution, by propagating a taint label along the derivation and checking it at the sink. But the taint formulation is monitorable only on a trace enriched to carry the derivation

label, which is the relation, and the typing of source and sink, which is the typing. Whether the barrier is treated as a hyperproperty requiring relational evidence or as a taint property requiring an enriched single trace, the unenriched action log cannot answer it, and what closes the gap in both cases is the same: provenance and legal-semantic typing added to the record. This is the criterion, instantiated.

Table 4. The information-barrier scenario, by governance layer.

Layer	What it observes	Can it answer "did the barrier hold"?
Per-action access policy	One agent, one resource, one action at a time	No. Each action is permitted; the policy never sees the pair
Single-path runtime monitor	The full ordered path of one agent	No. The breach is a relation between two paths; neither path contains it
Bare behavioural trace	That A read the deal room and B sent to the desk, both permitted	No. The trace records actions, not whether information flowed; the operative fact is not a function of the trace
Representation with typing and provenance	The read typed as protected-information-bearing; the send typed as barrier-crossing-capable; a provenance relation linking the send to the read	In principle yes, where the derivation is modelled: the relation is represented and can be checked

One might object that a sufficiently instrumented trace, carrying timestamps, lineage tags, access metadata, and cross-agent correlation, would catch the breach without any cybernetic apparatus. That is correct, and it is the criterion, not a refutation of it: a trace instrumented to those categories is no longer a bare trace but a provenance-bearing, relationally typed record, which is exactly what the criterion requires. The disagreement is over the name. The criterion does not require a particular ontology; it requires that the typing and the relation be represented somewhere in the supervisory evidentiary system, and a record lacking them cannot answer the determination however tamper-evident it is.

A published supervisory-evidence ontology for agentic AI under EU law supplies these primitives as a worked instance: a supervisory-evidence layer of conceptual slots, reuse of PROV-O for provenance and of the Data Privacy Vocabulary for protected-data classes, and SHACL shapes that make structural conditions such as delegation validity and revocation propagation checkable rather than asserted (Janssen, 2026b). That deposit is cited here as an existence proof that a representation satisfying the criterion can be built, and as the source of the case material for the experiment in Section 9, not as a validated reference architecture; it states its own limits, including a single-coder empirical base and pending validation, and the relational information-barrier shape used above is a natural extension of it rather than a feature of its current release.

The example also marks the limit of the criterion, which is consequential and is not minimised. The criterion closes the gap only for information flows that cross mediated, instrumented surfaces capable of provenance capture. A material and, on present evidence, uncharacterised fraction of consequential flow in real agent systems does not

cross such a surface; the paper does not quantify the fraction and does not have a basis to. Information can move through model-internal latent state, through tool calls that execute opaque code beneath the mediated interface, through generated code, through external APIs, through memory stores and retrieval systems, through summarisation that destroys provenance, through a human who reads one screen and types into another, and through timing and other side channels. None of these is captured by a representation over mediated events, and where the derivation runs through one of them the relation is not recorded and the breach is not answerable, by the representation or by anything built on it. This is a high practical limitation, stated as one. The criterion is confined to flows that cross provenance-capturable surfaces; for those flows a record without typing and relation cannot answer the determination and a record with them can. Extending the boundary is an instrumentation problem the criterion presupposes and does not solve.

7. The EU-Law Instance

The criterion is stated in domain-neutral terms. Its legal force, in this instance, comes from the obligations the EU AI Act places on high-risk systems, and the claim must be split across the Articles, in conditional terms, rather than asserted of all of them at once.

Article 12 requires that a high-risk AI system technically allow the automatic recording of events over its lifetime, for traceability. Article 14 requires that such a system be designed so that natural persons can effectively oversee it, which includes the ability to understand its operation, to intervene, and to override. Article 19 governs retention of the logs. Article 26 places on the deployer the duty to monitor operation and to keep generated logs. Article 72 requires a post-market monitoring system that systematically collects and analyses performance data. These obligations describe, in legal language, a control loop: a duty to sense, a duty to be able to decide, and a duty to be able to intervene.

The bearing of the criterion must be split across the Articles, because they demand different things. Article 12 is a recording duty. An uninterpretable log may well satisfy a substantial part of it: events are recorded, traceability of a kind exists, and the Article does not on its face require that the record be semantically interpretable by a supervisor. The paper does not claim that a bare trace fails Article 12. Article 14 is the stronger anchor. It requires that natural persons be able to effectively oversee the system, which includes understanding its operation well enough to intervene and override. Oversight that cannot answer the operative determination is not effective oversight. It is here, and at the Article 26 deployer duty to monitor, that the sensing failure bites. A record that satisfies the letter of Article 12 recording can still fail to support Article 14 oversight, and the gap between the two is exactly the typing and the relation the criterion identifies.

The claim is therefore stated conditionally, and as sufficiency in one direction and necessity in the other. A semantically and relationally structured supervisory record *can serve as evidence input toward* the Article 14 oversight obligation and the Article 26 monitoring duty; it is not, on its own, a discharge of them, which depends on facts about the deployment that no representation can supply. This is the sufficiency-direction claim, and it is held weak deliberately: the Act does not mandate a particular ontology, and the

criterion does not derive a mandated form from the statute. The converse is the stronger and more useful claim, and it is a necessity claim: a record that is not semantically and relationally interpretable *cannot serve as evidence that effective oversight under Article 14 was possible*, because the operative fact is not a function of the record's content, and a supervisor cannot read from the record what the Article requires to be establishable. A legal analysis of agents under EU law reaches a compatible conclusion from the doctrinal side, holding that high-risk agentic systems whose behaviour cannot be traced cannot satisfy the Act's essential requirements (Nannini et al., 2026). Untraceable, in the vocabulary of this paper, is unanswerable, and an oversight loop that cannot answer its determination cannot close. The point is confined to the oversight obligation; it is not extended to the recording obligation, which a bare log may meet.

Timing, in conditional terms, perishable. Under Article 113 of Regulation (EU) 2024/1689 as enacted, the high-risk obligations, including Articles 12 and 14, become applicable on 2 August 2026 for standalone Annex III systems and 2 August 2027 for high-risk systems embedded in regulated Annex I products. The Commission's Digital Omnibus on AI, a proposal adopted in November 2025, would defer these. A political agreement on the Omnibus on AI was reached in May 2026, and the agreed deferral moves the standalone Annex III date to 2 December 2027 and the Annex I product-embedded date to 2 August 2028. The mechanism of the deferral is material to this paper's argument: the Commission's stated rationale is to align the deadline for applying the high-risk obligations with the availability of related harmonised standards and support tools (European Commission, 2026). The obligations are being deferred precisely because the standards that would specify an oversightable record do not yet exist, which is the gap Section 11.2 identifies. As of this writing the agreement is a political agreement on a proposed amending regulation; it is not yet adopted and not yet published in the Official Journal, and until it is, the enacted dates of 2 August 2026 and 2 August 2027 remain the operative compliance dates. A reader should verify the current status against the Official Journal, as the dates may have moved by the time of reading. The substance of the obligations the argument concerns is unchanged by the deferral; only the dates move.

8. Two Models of Oversight, and What the Criterion Predicts

The criterion separates two models of oversight cleanly, and the separation is descriptive before it is normative.

Model 1, current oversight. A static or periodic audit produces a tamper-evident record of actions and decisions. The record has integrity. It is read by a human against a checklist. Relational and provenance facts are not represented; they are reconstructed, if at all, by manual inference after the fact.

Model 2, semantic oversight. A runtime sensing channel produces a record that types events to legal categories and carries the relations on which determinations depend. The record has integrity and answerability. A determination is recovered by a procedure over the record, not reconstructed by manual inference.

Table 5. The two models compared.

Property	Model 1, current	Model 2, semantic
Integrity of record	Yes, where hash-chained	Yes
Answerability of in-scope determinations	No, for relational and provenance facts	Yes, for flows crossing provenance-capturable surfaces
Interpretability by supervisor	Manual reconstruction, error-prone	By procedure over the record
Monitorability of relational properties	No, single-path or after-the-fact	With gray-box instrumentation and the relation represented
Legal sufficiency for Article 14 evidence	Cannot evidence that oversight was possible	Can serve as evidence input, not a discharge
False-finding risk	High, from manual inference over incomplete records	Lower for represented facts; capture risk remains (Section 10.6)
Runtime cost	Low	Higher, from instrumentation and representation

The separation licenses two falsifiable predictions, stated so that a reader can attempt to refute them.

Prediction 1. Organisations relying only on action logs without typing and relation will systematically fail to answer the in-scope Article 14 determinations, while organisations whose records carry the typing and relation will answer them, for flows crossing provenance-capturable surfaces. The experiment of Section 9 is the direct test.

Prediction 2. Increasing logging volume without adding typing and relation produces diminishing oversight value: the marginal determination answered per unit of log grows tends toward zero, because the determinations that remain unanswered are unanswerable in principle from records of that structure, not for want of volume.

Both predictions are falsifiable. Prediction 1 fails if a group given only bare logs answers the relational determinations as well as a group given the typed relational representation. Prediction 2 fails if increasing log volume, holding structure fixed, raises the proportion of in-scope determinations answered.

9. Empirical Validation: The Determination-Answerability Experiment

The criterion is a necessity claim about whether a determination is recoverable from a record. Such a claim is testable, and testing it is the difference between an argument and a result. It is also the only defence against a triviality reading: if real deployed logging already carries the typing and the relation, the criterion is trivial; if it systematically does not, and the determinations systematically fail to be answered from it, the criterion has bite. The experiment is designed to discriminate these cases. It is pre-registered; results are pending and are not reported here. The protocol is in Appendix B.

Design. A between-groups design with three independent expert panels, to avoid the learning-effect confound that arises when one panel sees all three conditions in sequence. Each panel receives the same set of cases, drawn from the 25-case EU automated-decision-making enforcement sample in the supervisory-evidence deposit

(Janssen, 2026b), reused here rather than constructed anew. Each case poses the same in-scope determinations: did protected data cross a boundary, was a human able to intervene, did the barrier hold, was the delegated authority valid.

- **Group A** receives an ordinary action log: events, timestamps, permit decisions, no typing, no relation.

- **Group B** receives the action log plus a provenance graph: the relation, without legal typing of events to categories.

- **Group C** receives the typed relational representation: typing and relation both present.

Measures. For each determination and panel: the proportion of determinations resolved correctly against a ground truth established by the case construction; the median time to resolve; inter-rater agreement within panel (Fleiss kappa); and self-reported confidence. The ground-truth coding of the cases is done by two coders, with inter-rater agreement reported, repairing the single-coder limitation that the underlying deposit carries.

Predicted result. Group A near the floor on the relational determinations (barrier, intervention window), because the operative fact is not in the record; Group C substantially above floor; Group B intermediate, able to answer the derivation-dependent determinations but not those requiring legal typing of categories. The non-relational determination that depends only on typing, not relation, separates Group B from Group C and is the within-design check on the minimality claim of Section 2.6.

The falsifiable core, stated as the triviality guard. The criterion predicts that Group A fails the relational determinations. If Group A answers them as well as Group C, the criterion is false or trivial, and the paper's central claim does not survive. The experiment is constructed so that this outcome is possible. That is what makes it a test rather than a demonstration.

Status. The design is pre-registered as Appendix B; the case sample, the SHACL validator, and the scenario files are available in the deposit. Recruitment, the power analysis fixing panel size, and the run are pending. The paper's argument does not presuppose the result; it states what the result would have to be for the criterion to hold, and exposes the criterion to that test.

10. Critical Evaluation

An argument of this kind is strengthened by stating its limits plainly, and by recording where the strongest available attacks land.

10.1 The criterion is necessity, not sufficiency

The central claim is that the typing and the relation are necessary for a record to answer an in-scope determination. It is not the claim that they are sufficient for good oversight. A supervisory system whose records satisfy the criterion can still lack a defined essential variable, an intervention repertoire that can act, a decision rule fit to the standard, a feedback latency shorter than the time in which harm sets, competent supervisors, or good faith, and can oversee badly with all the evidence it needs. The contribution identifies a precondition that the current emphasis on enforcement and tamper-evidence

omits. It does not supply a complete theory of oversight, and the paper should not be read at any point as claiming that the representation closes the loop on its own.

10.2 The cybernetic corroboration reaches only the coupled loop

The Good Regulator Theorem and the requisite-variety reading of Sections 4.1 and 4.2 apply only to a supervisory system coupled to the agent's future behaviour through an intervention repertoire. They do not reach a purely retrospective audit, which is an evaluator and not a regulator. The paper builds this into the argument rather than around it: where oversight is purely retrospective, the cybernetic corroboration is silent, and the criterion of Section 2, which is an evidentiary claim and does not depend on any cybernetic theorem, carries the case alone. This is why the contribution is placed in Section 2 and the cybernetics in Section 4, and not the reverse. If the cybernetic readings were withdrawn entirely, the criterion would stand on its construction and minimality proofs.

10.3 The variety reading is an application, and the unification with monitorability is conjecture

The requisite-variety reading is a qualitative application of Ashby's law, not a derivation from first principles within information theory, and the paper does not describe it as a proof. The stronger claim, that the variety deficit of static oversight and the single-trace monitorability limit are the same constraint, is conjecture. Touchette and Lloyd (2000) and Aguirre's channel-capacity formalisation (2025, Appendix A.5) are possible routes to formalising it, not an established bridge, and the paper does not rely on the unification. If it were established it would make the variety reading quantitative; if it failed, the two limits would remain two limits that point the same way, and the criterion would be untouched, since the criterion does not depend on the unification.

10.4 The hyperproperty classification is one formulation, and the monitorability limit is a black-box limit

The information-barrier breach in its strongest form is a 2-safety hyperproperty; in its operational form it is a taint property monitorable on a single enriched trace. The paper does not claim the breach is precisely noninterference; it claims that under both formulations the unenriched trace fails and the rescuing enrichment is the same typing and relation. The monitorability limit is a black-box limit, and "in general" is load-bearing: gray-box monitoring lifts it when the monitor is given knowledge of the system (Stucki et al., 2019), which is the criterion restated in the security vocabulary. Both qualifications support the criterion rather than undermining it.

10.5 Bounded completeness, not universal completeness

The two-feature set is claimed adequate only for the class of determinations defined by the selection rule in Section 2.2, and only where "relation" is read to carry provenance, authority, and temporal structure. The paper does not claim the set answers every determination one might pose, and does not claim that no further feature could be needed for an in-scope determination not yet considered. Universal completeness is the overreach the paper declines, on the same ground it declines an if-and-only-if theorem in 10.7.

10.6 A representation can be captured

Power (2007) describes the pathology by which formal procedures of risk management proliferate and satisfy auditors while displacing the substantive assessment they nominally represent. A semantic representation is not immune. It can be populated to produce the appearance of answerability while omitting the categories that would reveal a breach, in the same way a checklist can be completed without testing what matters. The criterion buys legibility, not truthfulness. It makes a record readable for a determination; it does not make the record honest. A legible record can be a curated one. The criterion raises the floor of what a supervisor can read, which is a real gain over a record that cannot be read at all, and it relocates the failure: with a bare trace the supervisor cannot see; with a gamed representation the supervisor can see what someone chose to show, which is auditable in a way the first is not. The gap between legibility and truthfulness is a governance problem the criterion lets one interrogate, not one it solves.

10.7 What the paper deliberately does not claim

For the avoidance of the overreach that would invite a triviality objection, the paper states three non-claims. It does not state an if-and-only-if completeness theorem, because the if-and-only-if would assert sufficiency, which 10.1 and 10.6 show to be false. It does not claim uniqueness of any representation, because Section 2.9 concedes functional equivalence among many representations, and uniqueness and functional equivalence cannot both hold. It does not promote the criterion to a theory of knowledge or a characterisation of the conditions under which knowledge can exist; the claim is the narrower one that a specific representation can or cannot answer a specific class of legal findings of fact.

10.8 The supervisory-evidence deposit is an existence proof, not a validated instrument

The representation invoked in Section 6 is published and accompanied by machine-checkable artifacts (Janssen, 2026b). It is explicit about its limits: it does not claim reference-architecture status, its empirical base is a single-coder reading of a case sample, and its validation tracks remain pending. The argument of this paper does not rest on the deposit's validation. It rests on the criterion, and the deposit serves as an existence proof that a representation satisfying the criterion can be built and as the source of the experiment's case material. The experiment in Section 9 repairs the single-coder limitation for its own ground-truth coding by using two coders and reporting agreement.

10.9 Strongest attacks, recorded

Three attacks deserve to be recorded because they are the ones a hostile reviewer would mount, and the paper's response to each is its narrowing rather than its defence.

The first is that the criterion is trivial: to evidence a relational legal predicate one must obviously record the relata, their types, and the relation. The response is that triviality is an empirical question, not a logical one, and is the precise thing the experiment tests. If deployed records already carry these features, the criterion is trivial and the experiment

will show Group A answering the determinations. If they do not, the criterion has bite. The paper does not assert non-triviality; it exposes it to test.

The second is that the cybernetic inversion is not novel. The response is concession: it is not, Aguirre (2025) made it, and the paper says so and locates its contribution elsewhere, in the criterion, the legal instance, the gap analysis, and the experiment.

The third is that Article 14 does not mandate a semantic representation. The response is concession on the sufficiency direction and insistence on the necessity direction: the Act mandates no ontology, the criterion derives none from the statute, and the operative claim is the converse, that a record without the structure cannot evidence that effective oversight was possible.

11. Implications

11.1 For boards and fiduciary accountability

A board that approves a high-risk AI deployment relies, for its account of control, on the records the deployment produces. If those records are behavioural traces without typing and relation, the board possesses integrity without answerability: it can show that a log was not altered and cannot show what the log means for the obligations it bears. The criterion gives that gap a name and a test. It is a sensing failure of the oversight loop, and it is the condition under which a determination is not a function of the record the board holds.

11.2 For the operationalisation of the EU AI Act

The Act requires logging and oversight without fully specifying the representation that would make a log oversightable, and at the time of writing the harmonised standards that would supply that specification are still in development. The Commission's own stated rationale for deferring the high-risk obligations is to align the deadline with the availability of those standards and support tools (European Commission, 2026). The criterion names the specification the standards presuppose and do not yet supply: a record that satisfies Article 12 recording can fail the Article 14 oversight the recording is meant to serve, and the typing and relation are the part the standard must specify for the recording to become oversightable.

11.3 For the design of oversight

Oversight of agentic systems should be designed as a closed control loop with a sensing channel that answers the determination, not as a periodic audit. Sensing requires a representation carrying typing and relation. Decision requires the ability to evaluate relational properties, which a single-path view cannot supply. Intervention requires a mechanism on the path that acts before effect. The static audit supplies none of the three at the structure the task demands. This is consistent with the double-loop conception of learning (Argyris and Schön, 1978): a static audit corrects deviations within a fixed frame, while oversight of a system that composes its own behaviour requires the governing frame itself to be testable, which a closed loop with a representation permits and a checklist does not.

11.4 Beyond AI governance

The criterion is stated over determinations, evidence, and answerability, none of which is specific to AI. Wherever a legal finding about an autonomous system depends on the system's execution path, on the relation between actions, or on the provenance of a flow, the criterion applies, and the failure it names recurs: a record can have integrity and still fail to answer the finding. Autonomous trading systems must answer whether a barrier held or an authority was valid; autonomous medical systems must answer whether a contraindication was visible at the decision point; critical-infrastructure controllers must answer whether a command derived from an authenticated source. Each is a finding of fact about events and relations, in scope by the rule in Section 2.2, and each is unanswerable from a bare event log for the same structural reason. The EU AI Act is the worked instance in this paper because its obligations are explicit and current. The criterion is stated so that the instance can be lifted to those domains, and demonstrated only for the instance worked here.

12. Conclusion

The question this paper asks is prior to whether oversight succeeds. It asks whether a runtime record can answer a legally operative determination at all. For a defined class of determinations, those that are findings of fact about specific events and their relations, the answer is a criterion: the record can answer the determination only if it carries the legal typing that maps events to the operative category and the relation on which the determination's truth depends. The criterion is a necessity claim, shown by construction for each determination in the class and shown to rest on a minimal property set, adequate for the class and not claimed beyond it. Sufficiency is disclaimed and shown to fail. The criterion is argued from the determinations, not from any borrowed theorem, and it is corroborated, not derived, by three results that converge on the same necessity from cybernetics and computer science, none of which the paper claims as novel, and one of which, the inversion of model-dependence onto the overseer, it attributes squarely to prior work.

Instantiated in EU AI Act oversight, the criterion yields a conditional legal claim: a record that is not semantically and relationally interpretable cannot serve as evidence that effective oversight under Article 14 was possible, even where it satisfies the Article 12 duty to log. The two duties come apart exactly at the typing and the relation. A gap analysis shows that no governance framework in current use produces, by itself, a record that satisfies the criterion, which is why the supervisory-evidence representation is the instrument the standards presuppose and do not yet specify. The criterion is testable, and the paper pre-registers the test rather than assuming its outcome.

The defensible form of the thesis, and the one the paper commits to, is this. For findings of fact about the execution paths of autonomous systems, a record is evidence of the finding only when it carries the typing and the relation; a record lacking them can raise a suspicion but cannot establish the finding, however tamper-evident it is. That claim is narrower than an impossibility result about static oversight, and it is the claim the argument supports, the gap analysis isolates, and the experiment tests. The structure is not specific to AI; EU AI Act oversight is where it is worked.

References

Aguirre, A. (2025) *Control Inversion: Why the Superintelligent AI Agents We Are Racing to Create Would Absorb Power, Not Grant It.* Future of Life Institute. Available at: https://control-inversion.ai/

Anderson, J. P. (1972) *Computer Security Technology Planning Study.* ESD-TR-73-51. Bedford, MA: United States Air Force Electronic Systems Division.

Argyris, C. and Schön, D. A. (1978) *Organizational Learning: A Theory of Action Perspective.* Reading, MA: Addison-Wesley.

Ashby, W. R. (1956) *An Introduction to Cybernetics.* London: Chapman & Hall.

Beer, S. (1979) *The Heart of Enterprise.* Chichester: John Wiley & Sons.

Chin, Z. S., Chiodo, M., Müller, D. and Snell, C. (2026) 'Reframing AI Loss of Control.' Preprint.

Chiodo, M., Müller, D., Siewert, P., Wetherall, J.-L., Yasmine, Z. and Burden, J. (2026) 'Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility.' Preprint.

Clarkson, M. R. and Schneider, F. B. (2010) 'Hyperproperties.' *Journal of Computer Security,* 18(6), pp. 1157–1210.

Conant, R. C. and Ashby, W. R. (1970) 'Every Good Regulator of a System Must Be a Model of That System.' *International Journal of Systems Science,* 1(2), pp. 89–97.

Espejo, R. (2001) 'Auditing as a Trust Creation Process.' *Systemic Practice and Action Research,* 14(2), pp. 215–236.

European Commission (2024) *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* OJ L, 2024/1689.

European Commission (2026) *An agile Digital Rulebook for the EU* and *Digital Omnibus on AI.* Shaping Europe's Digital Future. Available at: https://digital-strategy.ec.europa.eu/en/policies/digital-rulebook (last updated 7 May 2026; political agreement on the Omnibus on AI reached May 2026; not yet adopted or published in the Official Journal as of writing).

Janssen, J. (2026a) 'From Battlefield to Boardroom: Strategic Red Teaming as an Epistemic Governance Instrument in the Age of AI.' Working paper. SSRN.

Janssen, J. (2026b) 'A Supervisory-Evidence Ontology for Agentic AI under EU Law: Candidate Minimum Conceptual Set and Temporal Extension.' Working paper. Zenodo. DOI: 10.5281/zenodo.19758441.

Nannini, L., Leon Smith, A., Maggini, M. J., Panai, E., Feliciano, S., Tiulkanov, A., Maran, E., Gealy, J. and Bisconti, P. (2026) 'AI Agents Under EU Law: A Compliance Architecture for AI Providers.' Preprint, arXiv:2604.04604.

Power, M. (2007) *Organized Uncertainty: Designing a World of Risk Management.* Oxford: Oxford University Press.

Saltzer, J. H. and Schroeder, M. D. (1975) 'The Protection of Information in Computer Systems.' *Proceedings of the IEEE,* 63(9), pp. 1278–1308.

Schneider, F. B. (2000) 'Enforceable Security Policies.' *ACM Transactions on Information and System Security,* 3(1), pp. 30–50.

Stucki, S., Sánchez, C., Schneider, G. and Bonakdarpour, B. (2019) 'Gray-Box Monitoring of Hyperproperties.' In: ter Beek, M., McIver, A. and Oliveira, J. (eds.) *Formal Methods – The Next 30 Years (FM 2019).* LNCS, vol. 11800. Cham: Springer, pp. 406–424. DOI: 10.1007/978-3-030-30942-8_25.

Thobani, I. (2024) 'A Triviality Worry for the Internal Model Principle.' *Synthese,* 204(1), article 36. DOI: 10.1007/s11229-024-04693-x.

Touchette, H. and Lloyd, S. (2000) 'Information-Theoretic Limits of Control.' *Physical Review Letters,* 84(6), pp. 1156–1159.

Virgo, N., Biehl, M., Baltieri, M. and Capucci, M. (2025) 'A "Good Regulator Theorem" for Embodied Agents.' Preprint, arXiv:2508.06326 (ALIFE 2025).

Wang, C. L., Singhal, T., Kelkar, A. and Tuo, J. (2025) 'MI9: An Integrated Runtime Governance Framework for Agentic AI.' Preprint, arXiv:2508.03858.

Wang, H., Poskitt, C. M. and Sun, J. (2025) 'AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents.' Preprint, arXiv:2503.18666.

Cite this paper

```
@techreport{janssen2026record,   author      = {Janssen, Jeroen},   title       = {From Record to
Finding: Why Tamper-Proof Logs Cannot Establish Legal Oversight of Agentic AI},   institution =
{Apparens},   year        = {2026},   month       = {June},   type        = {Working paper},
doi         = {10.5281/zenodo.21025237},   note        = {Licensed under CC BY 4.0} }
```

Appendix A. Claim Inventory and Falsification Register

This appendix decomposes the paper into discrete claims, grades each by type, load-bearing status, and evidential strength, states a falsification condition for the load-bearing claims, and records the result of the tests run against the most exposed of them. It is included for scientific transparency, so a reader can take any claim, read the condition under which it would be false, and attempt to falsify it; and for consistency, since a paper arguing that oversight requires evidence graded by strength should hold its own claims to that standard. The register directs adversarial testing at the claims whose failure would cost the thesis.

A.1 Grading legend

Claim type. Def = definitional. Int = interpretive. Inf = inferential. Emp = empirical. Leg = legal. Nov = novelty. Nrm = normative.

Load-bearing status. Critical = the thesis collapses or must be materially narrowed if the claim is false. Supporting = failure weakens but does not collapse. Peripheral = illustrative.

Evidential grade. A = canonical or established. B = well-supported. C = defensible inference, contestable. D = conjecture or proof-of-absence. E = unverified or perishable.

The test target is the intersection of Critical with C, D, or E. Grades below reflect the state after the tests in A.5.

A.2 Master register

ID	Claim (summary)	Type	Load	Grade
K1	An agentic system selects its execution path at run time; the path is not enumerable in advance and can be steered by content the agent reads	Def/Emp	Critical	A
K2	A determination D is a legally operative finding of fact; E answers D iff D's truth value is a function of E's content	Def	Critical	A
K3	The in-scope class is defined by a selection rule: binary findings of fact about specific events and their relations, not evaluative or statistical properties	Def	Critical	B
K4	For an in-scope D, E answers D only if E carries a typing to the legal category and the relation on which D's truth depends (the criterion)	Inf	Critical	B
K5	A representation carrying neither typing nor relation can raise a suspicion but not establish a finding (the risk-versus-finding line)	Inf	Critical	B
K6	Necessity holds by construction for each of the four determinations: the bare log does not make D's truth value recoverable	Inf	Critical	B
K7	The property set {typing, relation}	Inf	Critical	B

ID	Claim (summary)	Type	Load	Grade
	is minimal: dropping either loses some in-scope determination			
K8	The set is adequate for the in-scope class where "relation" carries provenance, authority, and temporal structure (bounded completeness)	Inf	Critical	C
K9	Sufficiency fails: a record satisfying the criterion can be captured and is not thereby honest	Inf	Critical	B
K10	The criterion is satisfied by structure, not by a particular schema; functional equivalence holds across many representations	Int/Inf	Critical	B
K11	No governance framework in current use produces, by itself, a record satisfying the criterion for the in-scope class	Emp/Inf	Critical	C
K12	Tamper-evidence establishes integrity but not answerability; eIDAS Art. 41(2) gives a qualified timestamp a presumption of integrity of bound data	Int/Leg	Supporting	B
K13	Requisite variety: a once-acting regulator is in the wrong class to hold a runtime-dependent variable in range,	Inf	Supporting	C

ID	Claim (summary)	Type	Load	Grade
	as the sole mechanism			
K14	The regulator mapping holds only where the legal condition is made measurable, sensed, decisionable, intervention-tied	Def/Inf	Supporting	C
K15	The Good Regulator Theorem supports model-dependence for a coupled regulator only, not for a retrospective evaluator	Int	Critical	B
K16	The theorem is type-agnostic; the legal content of the model is supplied by the criterion, not the theorem	Int	Critical	B
K17	Noninterference is a 2-safety hyperproperty; a single-path black-box monitor cannot evidence it, and gray-box knowledge of the system lifts the limit	Emp/formal	Supporting	A
K18	Schneider's EM class is contained in the safety properties	Emp/formal	Supporting	A
K19	The three borrowed results converge on the same necessity but are not one theorem (conjecture, disclosed)	Inf	Supporting	D
K20	The cybernetic inversion onto the overseer is not novel; Aguirre (2025) made it as an impossibility argument, with	Nov/Emp	Critical	B

ID	Claim (summary)	Type	Load	Grade
	the theorem-weaker-than-slogan observation and a channel-capacity formalisation			
K21	The contribution is the conjunction the prior work leaves open: the criterion, its EU-law instance, the gap analysis, and the test	Nov	Critical	C
K22	EU AI Act Art. 12 requires automatic recording; Art. 14 requires effective human oversight including understanding, intervention, override	Leg	Critical	A
K23	A bare log may satisfy Art. 12 recording while failing Art. 14 oversight	Leg/Inf	Critical	C
K24	A non-interpretable record cannot serve as evidence that effective Art. 14 oversight was possible (necessity direction); a structured record can serve as evidence input, not a discharge (sufficiency direction, weak)	Leg/Inf	Critical	C
K25	Default dates 2 Aug 2026 (Annex III) / 2 Aug 2027 (Annex I); Digital Omnibus political agreement May 2026 defers to 2 Dec 2027 / 2 Aug 2028, not yet adopted or in the OJ; deferral tied to standards availability	Leg/Emp	Supporting	B

ID	Claim (summary)	Type	Load	Grade
K26	Nannini et al. hold that untraceable agentic behaviour cannot satisfy the Act's essential requirements	Emp/cite	Supporting	B
K27	The criterion's coverage is bounded to flows crossing provenance-capturable surfaces; model-internal, opaque, retrieval, summarisation, and human-transfer channels are outside it	Emp/Inf	Critical	C
K28	The two-model comparison and the two predictions are falsifiable as stated	Inf	Supporting	B
K29	The experiment (between-groups, three panels) discriminates the criterion from triviality; Group A predicted to fail relational determinations	Emp	Critical	C
K30	The supervisory-evidence deposit is an existence proof supplying the primitives (PROV-O, DPV, SHACL), not a validated reference architecture	Emp	Supporting	A
K31	Beer's System 3* and Espejo (2001) treat audit as a regulatory/trust function but not over agentic traces, hyperproperties, or EU-law evidence	Int/Emp	Supporting	B
K32	The criterion is	Def	Supporting	B

ID	Claim (summary)	Type	Load	Grade
	not a theory of knowledge; it concerns whether a specific representation answers a specific class of findings			
K33	The structure generalises to runtime oversight of autonomous systems beyond AI; demonstrated only for the EU AI Act instance	Inf	Supporting	C

Counts: 33 claims. Critical: 18. At-risk critical set (grade C or D): K8, K11, K21, K23, K24, K27, K29. No critical claim is graded D; K19 (Supporting) is the only retained conjecture, disclosed as such.

A.3 Falsification conditions for the at-risk critical claims

Only the falsifier is hunted; searching for support is excluded by design.

K4, K6 (the criterion and its construction). Falsifier: an in-scope determination that a bare, untyped, provenance-free record answers on its own, that is, whose truth value is a function of an event log carrying neither typing nor relation. Test: per-determination logical construction, paired with the experiment (K29).

K7 (minimality). Falsifier: a demonstration that one of the two features is redundant, that is, that every in-scope determination answerable with both features is answerable with one. Test: construction.

K8 (bounded completeness). Falsifier: an in-scope determination, under the selection rule, that the two-feature set cannot answer even with "relation" read to carry provenance, authority, and temporal structure. Test: construction across the determination class; expert read.

K11 (gap analysis). Falsifier: a governance framework in current use that, by itself, produces a record typing events to legal categories and carrying the relations on which findings depend, for the in-scope class. Test: framework-by-framework examination.

K21 (novelty of the conjunction). Falsifier: a single publication combining the overseer inversion with the legal-supervisory-evidence criterion for agentic AI. Test: adversarial prior-art search across cybernetics, AI governance, runtime verification, information-flow, and evidence-law literatures.

K23, K24 (the Article split). Falsifier for K23: a reading on which Art. 12 already requires semantic interpretability, or on which Art. 14 is satisfied by a bare log plus human attention. Falsifier for K24: a reading on which a non-interpretable record can evidence that effective oversight was possible. Test: EU AI Act legal read.

K27 (coverage boundary). Falsifier: a demonstration that the named uninstrumented channels are in fact provenance-capturable by the representation. Test: construction; disclosed as a limit, not resolved.

K29 (the experiment discriminates). Falsifier: Group A answering the relational determinations as well as Group C, which would show the criterion trivial or false. Test: the pre-registered run (Appendix B).

A.4 Structural choices of the argument

The register reflects several deliberate choices about how the argument is built, each made to keep a borrowed result from carrying a conclusion it cannot bear.

The criterion (K4 to K11) is argued independently from the determinations and carries the thesis on its own; the cybernetic results (K13, K15, K17, K19, K20) are positioned as convergent corroboration rather than as the source of the criterion. This matters because the inversion of model-dependence onto the overseer is not novel: it is Aguirre's (2025), made as an impossibility argument, with the theorem-weaker-than-slogan observation and a channel-capacity formalisation already present, and with oversight explicitly bracketed by Aguirre as distinct from control (his footnote 3). The contribution is therefore located in the criterion, its legal instance, the gap analysis, and the experiment, not in the cybernetic move, which is disclaimed as prior.

The legal claim is held as two directions of unequal strength: a weak sufficiency direction and a strong necessity direction (K24). The hyperproperty material is stated as an instrumentation requirement rather than an impossibility, citing the gray-box result directly (K17). The selection rule (K3), the minimality argument (K7), the bounded-completeness argument (K8), the gap analysis (K11), the two falsifiable predictions (K28), and the pre-registered experiment (K29) are what carry the criterion from an assertion to a testable result. Three non-claims are stated to forestall overreach: no if-and-only-if theorem, no uniqueness, no theory of knowledge (K32).

A.5 Status log

Claim	Test	Outcome	Result
K4 / K6 / K7	Logical construction plus real-system search	Survived; graded B	The bare log fails to make each of the four determinations recoverable; dropping either feature loses a determination. The typing-plus-relation structure reappears under another name in every candidate that answers the determination (taint, lineage, policy graph, proof circuit), confirming functional equivalence (K10) rather than uniqueness.

Claim	Test	Outcome	Result
K8	Construction across the class	Survived, narrowed; graded C	The two temporal/authority determinations are answerable only if "relation" carries authority and temporal structure; the claim is held bounded to the class, not universal.
K11	Framework-by-framework examination	Survived; graded C	NIST and ISO/IEC 42001 govern process, not evidence; Art. 12 and generic logging produce records without typing and relation; provenance carries the relation but not the legal typing. No framework supplies both by itself. Expert read on the standards landscape is the appropriate confirmation.
K15 / K16	Literature hunt plus reading of the proof	Survived; narrowed	No extension of the Good Regulator Theorem to an uncoupled retrospective evaluator found; the coupling scope is necessary and stated. The proof supports model-dependence, not a usable-representation claim (Thobani, 2024; Virgo et al., 2025); the legal content is sourced from the criterion, not the theorem.
K20 / K21	Primary-source verification of Aguirre; adversarial prior-art search	Verified; novelty narrowed to the conjunction	Aguirre (2025) confirmed to apply requisite variety and the Good Regulator Theorem to the overseer, to observe the theorem is weaker than its slogan (his footnote 101), to build a channel-capacity formalisation (his Appendix A.5), and to bracket oversight as distinct from control (his footnote 3). The inversion is therefore prior and is

Claim	Test	Outcome	Result
			disclaimed. No publication combining the inversion with the legal-supervisory-evidence criterion for agentic AI was found; K21 held at C as a bounded proof-of-absence.
K17	Formal-methods reading	Verified	Noninterference is 2-safety; black-box single-path monitoring cannot evidence it; gray-box monitoring with system knowledge lifts the limit (Stucki et al., 2019). The instrumentation reading is the correct one; the blunt impossibility reading does not hold.
K25	Primary-source verification	Verified, perishable	Enacted dates 2 Aug 2026 / 2027 confirmed; Digital Omnibus political agreement May 2026 confirmed via the Commission digital-rulebook page; deferral tied to standards availability; not yet adopted or in the OJ as of writing. Stated conditionally; grade B, perishable.
K26	Primary-source verification	Verified	Nannini et al. confirmed; the operative holding is that untraceable behavioural drift cannot currently satisfy the Act's essential requirements.
K12	Primary-source verification	Verified	eIDAS Article 41(2) gives a qualified electronic timestamp a presumption of the accuracy of date and time and the integrity of the bound data.
K30	Primary-source verification	Verified	The Zenodo deposit (DOI 10.5281/zenodo.19758441) resolves and supplies the primitives claimed: 23 slots,

Claim	Test	Outcome	Result
			PROV-O and DPV reuse, SHACL shapes, temporal extension; it states its own single-coder and pending-validation limits.
K29	Pre-registered design	Pending	The between-groups experiment is designed and its case material is available; recruitment, power analysis, and run are pending. The design is constructed so Group A answering the relational determinations would falsify the criterion.
K8, K11, K23, K24	Expert reads	Pending	Bounded completeness, the gap analysis on the standards landscape, and the Article 12/14 split are provisionally held pending a formal-methods read and an EU AI Act legal read, which the paper flags rather than presumes.
K27, K33	Disclosed limits	Open, disclosed	The coverage boundary and the cross-domain generalisation are stated as limits and as a generalisation demonstrated only for the worked instance.

The most exposed claims survived attack and hold in narrow, precise form, with the nearest prior work named and the borrowed results positioned as corroboration. The remaining open items are disclosed limits, pending expert reads, and the pending experiment, not unexamined assumptions.

Appendix B. Experiment Protocol (Pre-Registered)

This appendix specifies the determination-answerability experiment of Section 9 in enough detail to pre-register and to replicate. Results are not reported; the protocol is fixed before the run.

Hypothesis. For in-scope determinations that depend on a relation (the information barrier; the intervention window), the proportion correctly resolved is ordered Group C > Group B > Group A, with Group A at or near the floor. For the determination that depends on typing but not relation (the protected-class boundary crossing on a directly

typed channel), the proportion is ordered Group C > Group A with Group B near Group A, isolating the typing feature.

Design. Between-groups, three independent expert panels, one condition each, to remove the learning-effect confound of a within-subjects sequence.

Materials. The 25-case EU automated-decision-making enforcement sample from Janssen (2026b), reused. Each case is rendered in three record conditions:

- Condition A: action log only (events, timestamps, permit decisions).

- Condition B: action log plus provenance graph (relation present, legal typing absent).

- Condition C: typed relational representation (typing and relation present), realised through the deposit's vocabulary and SHACL profiles.

Ground truth. Each case's determinations are coded to a ground-truth value by two independent coders, with disagreements adjudicated and inter-rater agreement (Cohen kappa) reported. This repairs the single-coder limitation of the underlying deposit for the experiment's own coding.

Participants. Practitioners with relevant expertise (data protection, AI assurance, audit). Panel size fixed by a power analysis targeting the predicted between-group difference on the relational determinations at conventional power; the analysis is run and the target n fixed before recruitment. Assignment to panels is randomised.

Procedure. Each participant receives the cases in their panel's condition and records, per determination: a resolution (yes / no / cannot determine from this record), time to resolve, and confidence (ordinal). "Cannot determine from this record" is an explicit and encouraged option, because it is the response the criterion predicts for unanswerable determinations and conflating it with a guess would mask the effect.

Measures. Primary: proportion of determinations resolved correctly against ground truth, by condition and determination type. Secondary: proportion answered "cannot determine," median time, within-panel agreement (Fleiss kappa), confidence.

Falsification. The criterion is falsified, or shown trivial, if Group A resolves the relational determinations correctly at a rate not distinguishable from Group C. The minimality claim is challenged if Group B resolves the typing-dependent determination at the rate of Group C. Both outcomes are possible under the design and are the registered kill conditions.

Analysis. Pre-specified: between-group comparison on the primary measure for each determination type, with correction for multiple determinations; the "cannot determine" rate as a manifest check that the floor effect in Group A is an answerability failure and not a competence failure.

Status. Design and materials fixed; recruitment, power analysis, and run pending. The paper's argument does not presuppose the outcome.