The Cardinal Decision Evidence Assessment v1.2

The Cardinal Decision Evidence Assessment

A published method for testing whether AI-assisted decisions can be defended after the fact

Ronke Jegede, LLB (Hons)

Cardinal AI Systems, London, United Kingdom

ORCID: 0009-0004-7492-6356

Version 1.2 · 19 October 2026 · Licence: CC BY 4.0

Abstract

Regulated firms may be unable to rebuild the AI-assisted decisions they have already made and few know whether they can, because the question is rarely asked until the decision is being contested. This document sets out an applied test for whether they can.

Six tests examine whether the inputs, system state and human judgement behind a specific past decision were captured, whether that record survives unaltered, whether it persists as long as the firm's liability and whether an outsider could use it. Results resolve through a strict ordered procedure to one of four tiers: Defensible, Partially Defensible, Attested Only and Opaque. The procedure is applied twice, once against the record as it stands today and once against the record that will still exist at the end of the firm's liability horizon, so that a firm which is defensible today and will not be when challenged is reported at both points. What survives to the second assessment is set by a per-component retention schedule, summarised by the Jegede Retention Ratio (R).

The document also identifies the Cardinal Retention Asymmetry: under Regulation (EU) 2024/1689, technical documentation must be retained for ten years (Art. 18), decision logs for at least six months (Arts. 19, 26) and the right to explanation of an individual decision carries no stated limit (Art. 86). A firm may comply with every retention floor and still be unable to answer the question when it is asked.

Keywords: AI governance; record-keeping; EU AI Act; Article 12; retention; decision evidence; audit; financial services; Cardinal Retention Asymmetry; Jegede Retention Ratio

0. Citation

Jegede, R. (2026). *The Cardinal Decision Evidence Assessment v1.2*. Cardinal AI Systems. DOI: 10.5281/zenodo.23479027. Available at: https://cardinalaisystems.com/assessment/

Short form: Cardinal Decision Evidence Assessment v1.2 (CDEA v1.2).

This version supersedes v1.1 (DOI: 10.5281/zenodo.21952103, 15 August 2026), which remains permanently available, as does v1.0 (DOI: 10.5281/zenodo.21922417, 13 August 2026).

The changes in this version are editorial. The classification procedure, tier definitions, scoring states, the Test 5 measure and default liability horizons are unchanged. A result produced under v1.1 remains valid

and need not be re-scored. A result produced under v1.0 should be re-scored, as set out in the version history below. Citations to the concept DOI resolve to the most recent version.

Licensed CC BY 4.0. This document may be reproduced, adapted and applied freely, in whole or in part, with attribution. Firms may self-assess against it without engaging Cardinal AI Systems and without notification.

1. The assertion

A firm can only defend an AI-assisted decision if it can rebuild, after the fact, the inputs, the system state and the human judgement that produced it.

You cannot reconstruct what you did not capture, or what you no longer hold.

Everything below is the operational form of that single claim.

The proposition is not that AI systems must be explainable in the technical sense, nor that firms must retain more data in general. It is narrower and more demanding: that a specific past decision, challenged months or years later, can be rebuilt to a standard a regulator, a court or an insurer would accept.

This Assessment anticipates that many firms cannot do this and that fewer know whether they can, because the question is rarely asked until the decision is already being contested. That expectation is stated here as a hypothesis the Assessment is designed to test, not as a finding.

1.1 What is new here

The evidentiary gap this Assessment addresses has been articulated independently by regulators, practitioners and researchers. Srivastava and Sah (2026) argue that "evidence sufficiency is a missing layer of AI governance" and formalise the problem as the AI Evaluability Gap. Their treatment is conceptual: it defines six properties of evaluable evidence and two classes of certification, but does not supply an applied test.

This Assessment is an applied test. It is offered as one operationalisation of a problem others have described, not as a discovery of it. Section 7 maps the tests below to the Srivastava and Sah properties.

Solozobov (2026) specifies the Decision Evidence Maturity Model (DEMM), a post-hoc method for single-principal agentic AI deployments. It classifies each of seven decision-evidence properties into sufficiency categories for a given governance question, using an automated reconstructor. It then aggregates the results into a five-level maturity rubric running from ad hoc to continuously attested. This Assessment differs in four respects: it applies to any AI-assisted decision in a regulated firm, not only agentic systems; its unit is a decision type and the record architecture behind it; it is scored at two points in time, today and at the end of the liability horizon, with retention measured through the Jegede Retention Ratio, which DEMM does not address; and its final test is a rebuild performed by a person outside the firm rather than by software. The two are complementary. DEMM asks whether the evidence is sufficient now. This Assessment also asks whether it will still exist when the question is asked.

One finding here is original. Section 2.3 identifies the **Cardinal Retention Asymmetry**: a firm may be required to retain documentation about a system for ten years, evidence of an individual decision for six months and face a right to explanation of that decision with no stated time limit. The measure of a firm's exposure to it is the **Jegede Retention Ratio (R)**, defined at Test 5 and reported with every classification.

2. Scope

2.1 In scope and unit of assessment

Any decision affecting a customer, counterparty, employee or regulatory obligation where an AI or machine-learning system materially contributed to the outcome. The Assessment examines the decision record, not the model.

Designed for regulated firms that must evidence individual decisions to a third party: financial services, legal services, insurance and healthcare.

Unit of assessment. A tier attaches to a decision type and the record architecture supporting it, not to a firm. A firm running three decision types may hold three different tiers. Statements of the form "this firm is Tier 2" are imprecise and should be read as shorthand for "this decision type, at this firm, is Tier 2".

A single rebuild establishes a ceiling, not a rate. One successful rebuild shows that reconstruction was possible for that decision; it does not show that it is possible for the population. Where a firm wishes to claim a tier for a decision type as a whole, the exercise at Annex B should be repeated on a random sample of decisions spanning the liability horizon and the tier reported is the worst obtained. Reporting the best obtained is the attestation failure this Assessment exists to identify.

2.2 Out of scope

This Assessment deliberately does not address:

Model performance, accuracy or bias. A decision may be fully rebuildable and still be wrong. Evidentiary sufficiency is a necessary condition for defensibility, not a sufficient one.

Ethical acceptability of the use case. Whether a firm should deploy AI for a given purpose is a prior question.

Data protection lawfulness. UK GDPR obligations run alongside and are not discharged by passing this Assessment.

Model interpretability. The Assessment is indifferent to whether the model is a linear regression or a frontier LLM. The question is what the firm retained, not what the model did internally.

Vendor assurance, except insofar as it constrains the firm's ability to rebuild (Test 2).

2.3 Relationship to existing frameworks

Complementary to ISO/IEC 42001 and the NIST AI RMF, both of which address the management system and the risk process rather than decision-level evidence.

On the EU AI Act. Article 12(1) of Regulation (EU) 2024/1689 provides that "high-risk AI systems shall technically allow for the automatic recording of events (logs) over the lifetime of the system." Article 12(2) directs that logging capability towards identifying risk situations, facilitating post-market monitoring and monitoring operation. It does not direct it towards the reconstruction of an individual decision. The phrase "post-hoc reconstruction standard" has attached itself to Article 12 in commentary. One example states that "the EU AI Act's Article 12 post-hoc reconstruction standard requires that every decision a high-risk AI system makes be surfaceable from the record alone, without relying on the memory or interpretation of any individual in the deploying organisation."[1]

The phrase does not appear in the Act. More importantly, the obligation described is not the one Article 12 creates. Surfaceability of an individual decision to a person affected by it is the subject of Article 86, which confers a right on that person, subject to the qualifications set out below. Article 12 imposes a logging capability on the system, directed at risk identification and monitoring. A firm that reads the two as one obligation is assuming a duty the text does not create and may be assuming a capability its logs do not deliver.

The Act does, however, contain the asymmetry this Assessment is built to expose. This document names it the **Cardinal Retention Asymmetry**, defined here as the structural gap between the period for which a firm must retain documentation about a system and the shorter period for which it must retain the evidence of an individual decision that system produced, where the right to challenge that decision is subject to no stated limit at all.

Obligation	Period	Source
Technical documentation retained	10 years from placing on the market or putting into service	Art. 18(1)
Automatically generated logs retained, provider	at least 6 months	Art. 19(1)
Automatically generated logs retained, deployer	at least 6 months	Art. 26(6)
Right to explanation of an individual decision	no stated limit	Art. 86(1)

Distinguished from compliance asymmetry. Finch and Butt (2025) develop compliance asymmetry as a conceptual lens for the structural frictions between regulatory ambition and institutional capacity: the gap between what the AI Act requires and what low-capacity actors can realistically implement. The Cardinal Retention Asymmetry is a different object. It concerns two retention periods within a single firm's own obligations and would exist even where every actor was fully resourced and fully willing. The two are complementary. A firm may face both at once and this Assessment is one response to Finch and Butt's call for lightweight, role-specific compliance instruments adoptable by low-capacity actors.

Finch, W.W. and Butt, M. (2025). Gaps in AI-Compliant Complementary Governance Frameworks' Suitability (for Low‑Capacity Actors), and Structural Asymmetries (in the Compliance Ecosystem): A Systematic Review. *J. Cybersecur. Priv.* 5(4), 101. DOI: 10.3390/jcp5040101.

The three periods do not run from the same point. The ten years of Article 18(1) runs from the system being placed on the market or put into service, not from the decision. The six months of Articles 19(1) and 26(6) runs from the log entry. The right at Article 86(1) attaches to the decision and has no stated end. The asymmetry is unaffected by this and in most cases widened by it, because a decision taken late in a system's deployed life sits further from the start of the ten-year documentation clock than from the start of its own six-month log clock.

The asymmetry is not confined to this instrument. Any regime that separates system-level documentation from decision-level evidence and sets their retention periods independently, will exhibit it. The EU AI Act is the clearest current example because all three periods are stated in one text.

Ten years of documentation about the system; six months of evidence about the decision. Article 86(1) gives an affected person the right to "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken", a right that may be exercised long after the record establishing those elements has expired.

That right is qualified in two ways and both should be stated. It is scoped to Annex III high-risk systems other than those under Annex III point 2 (critical infrastructure). It applies only where the decision

"produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights". Neither qualifier narrows it for the sectors this Assessment addresses: a declined credit application and an insurance pricing decision both sit squarely within it.

Note also that creditworthiness evaluation and life and health insurance pricing are classified as high-risk (Annex III, points 5(b) and 5(c)) and that the Act reaches deployers established outside the Union where the output is used within it (Art. 2(1)(c)).

Application dates. Regulation (EU) 2026/1744 of 8 July 2026 (the Digital Omnibus on AI), OJ L, 24.7.2026, ELI: data.europa.eu/eli/reg/2026/1744/oj, in force 27 July 2026, amended Article 113 so that Chapter III, Sections 1, 2 and 3, which contain Articles 12, 19 and 26, apply from 2 December 2027 for systems classified as high-risk under Annex III and from 2 August 2028 for those classified under Annex I. Firms have longer than the original timetable allowed. They do not have less to do and the retention asymmetry above is structural rather than a function of when the obligations commence.

Two transitional rules apply. High-risk systems placed on the market or put into service before the relevant application date fall outside the obligations until they are subject to significant changes in design (Art. 111(2)). Providers and deployers of high-risk systems intended to be used by public authorities must comply by 2 August 2030 regardless.

The Omnibus did not amend Articles 12, 18, 19, 26 or 86. The record-keeping obligation, both retention floors, the ten-year documentation period and the right to explanation stand as originally enacted.

2.4 Status

A private methodology published by Cardinal AI Systems. Not issued, endorsed, approved or recognised by the FCA, the ICO, the European Commission or any regulator or standards body. Regulatory material is cited as evidence that the underlying problem has been publicly articulated, not as authority for the method proposed here.

3. Definitions

Decision record. The complete set of artefacts retained by the firm in respect of a single decision.

Rebuild (of a decision). The reconstruction of a specific past decision, its inputs, the system state that processed them and any human judgement applied, such that a competent third party can independently verify how the outcome was reached.

Rebuild is not replay. A rebuild does not require re-executing the model. Re-execution is frequently impossible: models are non-deterministic, versions are deprecated and vendor endpoints are withdrawn. What is required is a documentary account of what the system received, what state it was in, what it produced and what a person did with it, sufficient for a competent third party to verify how the outcome was reached. A firm that cannot re-run its model can still pass every test in this Assessment.

Evidentiary sufficiency. The threshold at which a decision record permits a rebuild by a party with no institutional knowledge of the firm. Assessed against the external reader, never against the firm's own understanding of its systems.

Liability horizon. The period during which the firm remains exposed to challenge in respect of a decision: the longest of the applicable limitation period, regulatory record-keeping requirement, contractual obligation or realistic complaint window.

Challenge event. Any occasion requiring a rebuild: a regulatory information request, a complaint escalation, litigation disclosure, an insurance claim or an internal investigation.

3.1 Terms this Assessment does not use interchangeably

Term	Question it answers	Why it is not a rebuild
Explainability	Why does this model behave as it does, in general?	Concerns model behaviour across cases. Silent on this case.
Auditability	Do records exist and can they be inspected?	Concerns existence. Does not test sufficiency.
Traceability	Can this output be linked to a process?	Establishes provenance of the artefact, not the substance of the decision.

A firm can be fully explainable, fully auditable, fully traceable and entirely unable to rebuild a single decision.

3.2 Two adjacent literatures and how this differs from both

Reconstructability analysis (Klir and Cavallo, 1979; Zwick, 2004) is an established discipline in general systems theory concerning the inference of a system's structure from partial observations of its parts. This Assessment is unrelated to it: the object here is a single past event, not a system's structure and the method is documentary rather than statistical.

Reconstruction attacks in the privacy and differential-privacy literature describe the recovery of individual records from aggregate statistics, a harm to be prevented. The usage here is the inverse: the deliberate, authorised rebuilding of the firm's own decision from its own retained records.

4. The tests

Six tests. Tests 1, 2, 3 and 4 are scored **Met / Partial / Not met**. Test 5 uses a retention schedule and ratio rather than a score. Test 6 uses four states of its own, set out below. Each test requires evidence. A firm's assertion that a test is met is not evidence that it is met.

Ordered by dependency: failure at Test 1 makes Tests 3, 4 and 6 largely academic. Test 5 is independent of the others and is computed from documented retention periods rather than from the rebuild exercise.

Grading rule. Met: the evidence required is present and complete for the decision under assessment. **Partial:** some but not all of the required evidence is present and the missing part can be inferred or supplied from another retained record without recourse to a person. **Not met:** required evidence is absent, or recovering it depends on an individual's recollection, or the condition falls within that test's stated failure criteria.

Where a test's *Fails where* list is satisfied, the score is Not met, not Partial. Partial is for incompleteness, not for the named failure modes.

Test 1: Input capture

Are the exact inputs to the decision retained, or only the output?

The failure this Assessment expects to encounter most often. Firms routinely retain the model's output and the eventual decision while discarding the prompt, the customer data as it stood at that moment, the documents supplied and any retrieved context.

Evidence required: Verbatim inputs as submitted, retrievable for a named individual decision, including all documents and data passed to the system.

Fails where: Inputs are summarised rather than retained; inputs are recoverable only by re-querying a live database that has since changed; the prompt is templated and only the variables were stored, without the template version.

Test 2: System state

Is the state of the system at the point of decision recorded?

Evidence required: Model identifier and version; material configuration parameters; version of any prompt template, system instruction or rules layer; identity and version of any retrieval corpus consulted; vendor and endpoint where applicable.

Fails where: The firm records the vendor but not the version; the vendor updates or deprecates the model and the firm cannot establish which version served the decision; a retrieval corpus has been updated with no snapshot or index versioning.

Vendor dependency. Where version-level information cannot be obtained from the provider, this test cannot be met by effort. It is a procurement failure, not an operational one and should be escalated as such.

Test 3: Human intervention

Where a person accepted, modified or overrode the AI output, is that act recorded with its rationale?

Evidence required: Identity of the individual; action taken (accepted, modified, overridden or escalated); the output as presented to them; and their rationale, in their own words, where they departed from it or accepted it in a case flagged for review.

Fails where: The system records only the final decision, making acceptance indistinguishable from independent judgement; “reviewed by” is captured without content; the interface does not retain what the model actually produced.

This test carries disproportionate weight. A firm claiming meaningful human oversight but unable to evidence any specific instance of that oversight operating is in a materially worse position than a firm that never claimed it.

Test 4: Temporal integrity

Can the decision be rebuilt as it stood on the decision date, rather than as the system stands today?

Evidence required: Demonstration that the record is immutable or version-controlled and that the rebuild does not depend on any component that has since changed without snapshot.

Fails where: The rebuild requires querying current customer records; policies or rules have been updated in place; logs are mutable; the exercise produces today’s answer to yesterday’s question.

Test 5: Retention horizon

Do the records survive as long as the liability does?

Evidence required: A documented comparison of actual retention periods for each component of the decision record against the liability horizon for that decision type.

Measure. Express the result as the **Jegede Retention Ratio (R)**, defined here as the retention period of the shortest-lived component of the decision record, divided by the liability horizon. The worst component governs, because a record missing one element cannot be rebuilt. Test 5 does not produce a score that enters the ordered procedure. It produces two outputs: a per-component retention schedule and R as the headline figure summarising it. The schedule determines which components are deemed absent at t_H and therefore drives the second classification. R identifies the earliest point of failure and is the figure reported. Where R is below 1.0, both tiers are reported; where it is 1.0 or greater, one tier is reported. The general grading scale at section 4 does not apply to this test.

Report the governing component with the ratio. R alone is diagnostically inert. R stated with the component that produced it tells a remediation team where to go. Logs retained twelve months against a six-year complaint window are reported as **R = 0.17 (application logs)**.

Default liability horizons. R is only comparable between firms if the denominator is set consistently. Assessors should use the following defaults and depart from them only with a stated reason recorded in the assessment.

Decision type	Default horizon	Basis
Consumer credit, including affordability	6 years, longer where the three years from awareness limb applies	Limitation Act 1980 s.5; FOS jurisdiction of 6 years from the event or 3 years from awareness, whichever expires later
Retail investment and advice	6 years, longer where the three years from awareness limb applies	Limitation Act 1980 s.5; FOS jurisdiction; DISP complaint record obligations
Insurance pricing and underwriting	6 years	Limitation Act 1980 s.5
Employment decisions affecting existing employees	6 years for contractual claims	Limitation Act 1980 s.5. Note that discrimination claims under the Equality Act 2010 run on much shorter limits and do not extend the horizon
Recruitment and selection decisions	Set by policy, with a stated basis	No contract exists, so s.5 does not apply. Equality Act limits are short. The horizon is a retention policy decision the firm must justify
Public authority decisions	Unbounded where the decision continues to have effect	Judicial review and continuing effect. See the note below

Where more than one basis applies, the longest governs. Where a firm's own contractual or policy commitments exceed these, those govern. Regulatory record-keeping floors that fall below the applicable limitation period, such as the five year minimum in the retained MiFID rules, do not set the horizon and should not be used as the denominator.

Unbounded horizons. Where the liability horizon is unbounded, as it may be for a public authority decision that continues to have effect, R is not computed. The assessment reports *unbounded horizon* in place of a ratio, together with whether the firm holds a documented, dated commitment to review retention against continuing effect, with a named owner. The re-score at t_H proceeds as set out at section 5.1: every component with a finite retention period is deemed absent.

Where a decision type has no stated horizon at all, including recruitment and selection decisions where the firm has not set one, the unbounded procedure applies until the firm records a horizon and its basis.

Relationship to data minimisation. R is a floor-setting measure, not a maximand. Nothing in this Assessment supports retaining personal data longer than a lawful basis permits and Article 5(1)(e) of the UK GDPR is not displaced by anything here. Where retention aligned to the liability horizon is required, the purpose is ordinarily the establishment, exercise or defence of legal claims. The lawful basis is ordinarily legitimate interests or compliance with a legal obligation, applied to the minimum record that permits a rebuild rather than to the whole customer file. Where special category data is involved, the legal claims condition at Article 9(2)(f) of the UK GDPR also applies. A firm that cannot justify retaining the decision record for the liability horizon should record that conclusion and its reasons and should expect to be unable to defend a challenge arriving after the record expires. Both positions are defensible. Holding neither is not.

Indicators of a low ratio. Application logs retained 90 days against a six-year complaint window; vendor-side logs running on the vendor's schedule and not exported; components expiring on different schedules, leaving a partial record. The last is the most common and the least noticed, because each component looks adequately retained when considered alone.

This test warrants particular attention. Retention is commonly configured against operational or storage-cost logic, never against the liability horizon. The result is a record that is complete on day one and incomplete by the time it is needed.

Where a firm has configured retention to a regulatory floor, note that floors are not horizons. The EU AI Act requires providers and deployers alike to keep automatically generated logs "for a period appropriate to the intended purpose of the high-risk AI system, of at least six months, unless provided otherwise in applicable Union or national law" (Arts. 19(1) and 26(6)).

The full wording matters and is usually dropped. Six months is a floor, not a specification. The operative test in the Act is what is appropriate to the intended purpose and where the intended purpose is a decision that can be challenged for six years, an argument that six months is appropriate is difficult to sustain. A firm that retains for six months has complied with the floor. It has not necessarily complied with the obligation and it will still be unable to answer the question when it is asked.

Test 6: Third-party rebuild

Could a competent person outside the firm perform the rebuild from the record alone?

The integrating test. Tests 1 to 4 examine components; this examines whether they cohere into something an outsider can use.

Evidence required: A completed rebuild, performed by someone without institutional knowledge, working only from the retained record, producing a documented account of the decision.

Failed on attempt where: The rebuild requires a conversation with the engineer who built the system; field names are undocumented; the record is complete but unintelligible.

Firms should perform at least one live rebuild annually on a randomly selected decision and record the result. This is the single highest-value action available under this Assessment.

Scoring states. Test 6 does not use the general Met / Partial / Not met scale. It admits four states:

Met. A rebuild was attempted by a person without institutional knowledge and succeeded.
Partial. A rebuild was attempted and produced an incomplete but usable account.

Failed on attempt. A rebuild was attempted and could not be produced from the retained record.
Not attempted. No rebuild has been performed.

The distinction between the last two is deliberate and carries different weight in classification. *Not attempted* is an absence of evidence: it blocks Tier 1, because Tier 1 asserts that reconstruction has been demonstrated rather than that it is believed possible, but it does not demote. *Failed on attempt* is evidence of absence: it is direct proof that the record does not support a rebuild, it is stronger than any inference drawn from the component tests and it demotes to Tier 3 or below.

5. Classification

Classification is a strict ordered procedure, evaluated twice.

Tests 1, 2, 3, 4 and 6 ask whether a decision *can* be rebuilt. Test 5 asks whether it will *still* be rebuildable when the question is asked. Those are different questions and they are scored at different moments.

5.1 Two points of assessment

At t_0, the point of assessment. Score Tests 1, 2, 3, 4 and 6 against the record as it exists today. Test 5 is not used at t_0.

At t_H, the end of the liability horizon. Deem every component of the decision record whose retention period is shorter than the liability horizon to be absent. Re-score Tests 1, 2, 3 and 4 on the surviving record. Where the loss of an expired component makes a test unanswerable, that test is Not met at t_H.

Test 6 at t_H. A rebuild cannot be performed at a future date, so Test 6 is not re-performed. It carries forward its score from t_0, with one exception: where the rebuilder relied on a component that expires before t_H, Test 6 is scored Failed on attempt at t_H.

This is the strongest evidence in the instrument and it is available at no extra cost. The assessor knows which components the rebuilder actually reached for, because the rebuild was performed and observed. A rebuild that depended on a log expiring in twelve months is a rebuild that will not be possible in thirteen. That is a finding, not an inference and it should be recorded during the exercise at Annex B by noting which components the rebuilder used.

Test 3 at t_H. Oversight records commonly sit in the same store as the decision logs and expire with them. Where that is so, Test 3 is Not met at t_H and step 5 fires wherever human oversight is claimed as a control. This is the commonest route from Tier 1 at t_0 to Tier 3 at t_H and it deserves particular attention from any firm whose governance rests on a named individual's oversight of automated decisions.

Where the horizon is unbounded. Where the liability horizon is unbounded, deem absent every component of the decision record that has any finite retention period and re-score on what remains. A decision that continues to have effect indefinitely, supported by records that expire, will not be defensible at the moment it is challenged. The result reports *unbounded horizon* with the review commitment status in place of a ratio.

This produces a determinate outcome and the assessor should expect it rather than treat it as a scoring error. Where every finite component is deemed absent, any rebuild that relied on a retained record scores Failed on attempt at t_H and the decision type resolves to Tier 3 or Tier 4 at t_H in every case. That is the correct result. A decision with indefinite effect and finite records cannot be defensible at an indefinite future date and no configuration of retention periods changes that. The only mitigation available is the documented review commitment, which is why Test 5 is scored against it and the only alternative is for the firm to record a bounded horizon and state its basis.

Where R is 1.0 or greater, no component expires, Test 6 carries forward unchanged and the two tiers are identical.

5.2 The ordered procedure

Apply these rules at each point of assessment. Evaluate top to bottom and stop at the first that applies. Every possible score vector resolves to exactly one tier.

Step	Condition	Tier
1	Test 6 is Failed on attempt *and* any of Tests 1, 2, 4 is Not met	4: Opaque
2	Test 6 is Failed on attempt	3: Attested Only
3	Two or more of Tests 1, 2, 4 are Not met	4: Opaque
4	Any one of Tests 1, 2, 4 is Not met	3: Attested Only
5	Test 3 is Not met *and* human oversight is claimed as a control	3: Attested Only
6	All applicable tests are Met	1: Defensible
7	Any other combination	2: Partially Defensible

Tier	Definition
1: Defensible	Individual decisions can be rebuilt by an external party.
2: Partially Defensible	Rebuilding is possible but incomplete, effortful or dependent on institutional knowledge.
3: Attested Only	The firm can evidence that a process existed. It cannot evidence what happened in any specific case.
4: Opaque	No meaningful rebuild is possible.

5.3 Reporting

Report both tiers with the ratio and its governing component:

Tier 1 at t_0 → Tier 3 at $t_H \cdot R = 0.17$ (application logs)

Where R is 1.0 or greater, report the single tier and state $R \geq 1.0$.

Conformance. A result quoted without both tiers and the ratio is not a CDEA result and should not be described as one. A firm that is Tier 1 today and Tier 3 at the end of its liability horizon is not defensible; it is defensible for as long as nobody asks. Quoting the t_0 tier alone conceals precisely the failure this Assessment exists to surface and it is the misuse this instrument is most likely to attract. This rule cannot be enforced under a CC BY licence. It can be stated and stating it is what allows a reader to recognise a partial quotation for what it is.

On the second scoring. The re-score at t_H is not a prediction about the future. It is a statement about the record as currently configured: given the retention periods the firm has set today, this is what will remain when the challenge arrives. The firm can change that outcome and the cheapest way to change it is usually a retention setting rather than a governance programme.

Not applicable. Test 3 alone admits a score of Not applicable, where no human intervened in the decision at any point. Fully automated decisions have no oversight act to record and scoring them Not met would penalise a firm for the absence of a control it never claimed. Step 6 reads "all applicable tests" accordingly; step 5 continues to apply where oversight is claimed.

On Test 6 at steps 1 and 2. Failed on attempt is evidence of absence: direct proof that the record does not support a rebuild, stronger than any inference drawn from the component tests. Not attempted is an absence of evidence: it blocks Tier 1 at step 6 but does not demote. Partial likewise blocks Tier 1 and falls to step 7.

On Test 4 at step 4. A rebuild that returns current data for a past decision is not an incomplete rebuild. It is a confidently misleading one and it is more dangerous than a gap, because it produces an answer nobody has reason to doubt. A firm failing Test 4 cannot evidence what happened; it can only evidence what is true now. That is the Tier 3 condition.

6. Why this question is being asked now

At page 17 of *AI and the future of retail financial services* (The Mills Review), the Review states:

"In regulated activity, a firm may need to show why one customer was treated differently from another. A useful answer is not enough if the basis for it cannot be reconstructed."

The same page identifies the failure this Assessment tests for, "a decision that the firm cannot properly evidence" and on human oversight states that firms "need to decide where human approval is required, what reviewers should see and how challenge should be recorded." That sentence is the substance of Test 3.

Elsewhere in the Review, on contestability: "firms need records that show what happened; the decision logic must be sufficiently understood to explain the outcome." On redress: "Firms will need auditable records, clear complaints routes and enough understanding of AI outputs to explain decisions and mitigate harm."

The Review does not prescribe how a firm should establish whether its records meet that bar and does not address retention periods. This Assessment is one method for answering the first question; the second is addressed at Test 5.

Note that the Review makes recommendations for consideration. It is cited here as evidence that the problem has been publicly articulated, not as a statement of supervisory expectation.

AI and the future of retail financial services (The Mills Review) (2026), p.17.

7. Mapping to the AI Evaluability Gap

Srivastava and Sah (2026) identify six properties of evaluable evidence. This Assessment sits wholly within what they term Operational Certification and operationalises three of the six at decision level:

Evaluability property	Corresponding test	Note
Observability	Tests 1, 2	Their "logged inference inputs and outputs" is the substance of Test 1.
Verifiability	Tests 3, 6	Their standard, "self-attested evidence is not verifiable evidence", is the basis of Tier 3.

Evaluability property	Corresponding test	Note
Temporal validity	Tests 4, 5	Partial. See below.
Attributability	Not addressed	Concerns isolating the system's causal contribution to an outcome. A different question.
Intervenability	Not addressed	Concerns control over the system, not the record.
Calibration	Not addressed	Concerns whether stated confidence is warranted, not whether evidence exists.

On temporal validity. Srivastava and Sah treat temporal validity as the decay of evidence into irrelevance: an experiment run on a population that no longer exists is no longer current evidence. Tests 4 and 5 address a prior and cruder question, whether the record still exists and whether it can be read as it stood on the decision date. Evidence that has been deleted cannot decay. This Assessment therefore addresses a failure mode beneath the one they describe.

On the relationship generally. The paper is explicitly conceptual and identifies as future work the need to operationalise its properties: "the six properties are stated qualitatively. Operationalizing them requires quantitative metrics." This Assessment is a partial, applied answer to that, restricted to record sufficiency and to operational rather than investment decisions.

Srivastava, V. and Sah, T. (2026). The AI Evaluability Gap: The Missing Layer for Managing Risk and Sustaining Value. arXiv:2606.21015v1 [cs.AI].

8. Declarations

Funding. This research received no external funding.

Conflicts of interest. The author is the founder of Cardinal AI Systems, which offers paid assessments applying this method. The method is published in full under CC BY 4.0 and may be applied by any firm without engaging Cardinal AI Systems and without notification.

Data availability. No datasets were generated or analysed. The worked example at Annex A is synthetic.

Use of AI tools. Drafting and editing were assisted by large language models. All legal citations, quotations and page references were verified by the author against primary sources.

9. References

Equality Act 2010, c. 15. United Kingdom.

Finch, W.W. and Butt, M. (2025). Gaps in AI-Compliant Complementary Governance Frameworks' Suitability (for Low-Capacity Actors), and Structural Asymmetries (in the Compliance Ecosystem): A Systematic Review. *Journal of Cybersecurity and Privacy*, 5(4), 101. DOI: 10.3390/jcp5040101.

Financial Conduct Authority (2026). *AI and the future of retail financial services* (The Mills Review).

Financial Conduct Authority. Handbook, Dispute Resolution: Complaints (DISP), in particular DISP 1.9 (complaints record rule) and DISP 2.8 (was the complaint referred to the Financial Ombudsman Service in time).

International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system. ISO, Geneva.

Klir, G.J. and Cavallo, R.E. (1979). Reconstructability analysis of multi-dimensional relations: a theoretical basis for computer-aided determination of acceptable systems models. *International Journal of General Systems*, 5(3), 143–171.

Limitation Act 1980, c. 58. United Kingdom.

National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. United States Department of Commerce.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation), as retained in United Kingdom law.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L, 2024/1689, 12.7.2024. ELI: data.europa.eu/eli/reg/2024/1689/oj.

Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). OJ L, 24.7.2026. ELI: data.europa.eu/eli/reg/2026/1744/oj.

Solozobov, O. (2026). Decision Evidence Maturity Model for Agentic AI: A Property-Level Method Specification. arXiv:2605.04093.

Srivastava, V. and Sah, T. (2026). The AI Evaluability Gap: The Missing Layer for Managing Risk and Sustaining Value. arXiv:2606.21015v1 [cs.AI].

Zwick, M. (2004). An overview of reconstructability analysis. *Kybernetes*, 33(5/6), 877–905.

Annex A: Illustrative application

Synthetic worked example. Illustrative only; does not describe any actual firm.

Scenario. A mid-sized UK lender deploys an LLM-based tool to summarise applicant documentation and flag affordability concerns for underwriter review. An application is declined in March. Twenty months later the applicant complains that the decline rested on a mischaracterisation of their income.

At t_0, the point of assessment.

Test	Finding	Score
1: Input capture	Uploaded documents retained. Prompt template not versioned; only variable fields stored, a stated failure condition for this test.	Not met
2: System state	Vendor and model family recorded. Specific version not. Vendor updated the model in May.	Not met
3: Human intervention	Underwriter's decision recorded. The AI-generated	Not met

Test	Finding	Score
	summary as presented to them was not retained; no rationale captured.	
4: Temporal integrity	Applicant records updated since. Rebuild returns current data.	Not met
6: Third-party rebuild	Not attempted.	Not attempted

Step 1 does not apply because Test 6 was not attempted. Step 3 applies: Tests 1, 2 and 4 are all Not met, which satisfies the two or more condition. **Tier 4 at t_0.**

Test 5. Application logs are retained twelve months against a six-year complaint window. The shortest-lived component is the application log. **R = 0.17 (application logs).**

At t_H, the end of the liability horizon. The application logs have expired. Tests 1 and 2 were already Not met and remain so. The position does not improve. **Tier 4 at t_H.**

Classification: Tier 4 at t_0 → Tier 4 at t_H · R = 0.17 (application logs).

This firm is a rare case in which the two tiers agree, because its record was already insufficient before anything expired. The more common and more instructive pattern is a firm that is Tier 1 or Tier 2 at t_0 and falls to Tier 3 or Tier 4 at t_H. Such a firm is not badly governed today. It has simply configured retention against storage cost rather than against the period during which it can be asked to account for the decision.

The firm cannot establish what the model said, which model said it, what the underwriter saw or what the applicant's file contained at the time. It can establish only that a decline was recorded with an underwriter's name attached.

Nothing in this outcome required the model to have malfunctioned. The exposure arises entirely from the record and the record was configured before anyone considered the question.

Remediation sequence. Retention first (Test 5). It is a configuration change and is the cheapest fix with the largest effect across the liability horizon. Then Test 3, requiring an interface change to retain what the operator was shown. Then Test 2, a procurement conversation. Tests 1 and 4 follow from a versioned, immutable decision-record store.

Annex B: Self-assessment

1. Select a single decision type, not the whole estate.
2. Select one completed decision from at least nine months ago. That is long enough for common log expiries and vendor model updates to have taken effect, while still far short of a typical liability horizon.
3. Assign the rebuild to someone who did not build the system.
4. Give them the retained record and nothing else.
5. Ask them to produce an account of how the decision was reached.
6. Score Tests 1 to 4 on what the rebuilder needed and could not find. Score Test 6 on whether the rebuilder succeeded, failed on the attempt, or produced a partial account. Record which components of the decision record the rebuilder actually used, as this determines Test 6 at t_H. Classify. This is the tier at t_0.
7. Compute Test 5 separately from documented retention periods against the liability horizon. It does not depend on the rebuild exercise. Record the ratio and the governing component.

8. Identify every component of the decision record whose retention is shorter than the liability horizon. Where the horizon is unbounded, follow the procedure at section 5.1 instead and deem absent every component with any finite retention period. Deem those components absent and re-score Tests 1 to 4 on what remains. Carry Test 6 forward, scoring it Failed on attempt if the rebuilder used any component now deemed absent. Classify again. This is the tier at t_H.
9. Report both tiers with the ratio, its governing component and the decision type assessed.

A firm should expect to establish its tier for one decision type within a working day.

Standard result string

Results should be reported in a single consistent form so that they can be compared and cited:

CDEA v1.2 · Tier 1 at t_0 → Tier 3 at t_H · R = 0.17 (application logs) · consumer lending declines · assessed [YYYY-MM-DD]

Where the liability horizon is unbounded, the ratio is replaced by the horizon status and the review commitment:

CDEA v1.2 · Tier 2 at t_0 → Tier 4 at t_H · unbounded horizon (review commitment held) · planning enforcement decisions · assessed [YYYY-MM-DD]

The elements are: instrument and version; both tiers where R is below 1.0, or one tier where it is 1.0 or greater; the ratio with its governing component, or the horizon status and review commitment where the horizon is unbounded; the decision type assessed; and the assessment date.

Version history

Version	Date	Change
1.0	13 August 2026	Initial publication. DOI: 10.5281/zenodo.21922417.
1.1	15 August 2026	DOI: 10.5281/zenodo.21952103. Substantive revision. Classification outcomes have changed; results produced under version 1.0 should be re-scored. Changes are grouped below.
1.2	19 October 2026	DOI: 10.5281/zenodo.23479027. Editorial clarification. The classification procedure, tier definitions, scoring states, the Test 5 measure and default liability horizons are unchanged. Results produced under version 1.1 remain valid and need not be re-scored.

Changes in version 1.2

Data minimisation. The relationship to UK GDPR Article 5(1)(e) at Test 5 now distinguishes the purpose of retention (the establishment, exercise or defence of legal claims) from its lawful basis

(ordinarily legitimate interests or a legal obligation) and notes the separate condition for special category data. The previous wording treated the purpose as the route. No classification outcome is affected.

Related work. Section 1.1 now notes the Decision Evidence Maturity Model (Solozobov, 2026) and states how this Assessment differs from it. The reference is added at section 9.

Editorial. The citation at section 0 and the standard result strings at Annex B now refer to version 1.2. Punctuation has been brought into line with the author's house style. No wording that affects meaning has changed. The abstract uses the wording of the description on the version 1.1 record. The PDF deposited as version 1.1 carried an earlier draft of the abstract; the body of that document is the same.

Changes in version 1.1

Classification. The procedure is now time-indexed and evaluated at two points: t_0 (the point of assessment) and t_H (the end of the liability horizon). Both tiers are reported. The horizon-limited annotation used in version 1.0 is withdrawn, superseded by the two-tier report, which states the same thing more precisely. A conformance rule has been added: a result quoted without both tiers and the ratio is not a CDEA result. Section 5 has been renumbered and the retention ratio note that was section 5.1 in version 1.0 is now distributed across sections 5.1 and 5.3.

Test 5. Removed from the ordered procedure. It now produces a per-component retention schedule, which determines which components are deemed absent at t_H; and R as the headline figure summarising that schedule and identifying the earliest point of failure. A default liability horizon table has been added so that R is comparable between firms. The governing component is now reported with the ratio. Unbounded liability horizons are addressed at both points of assessment.

Test 6. Split into four scoring states, distinguishing a rebuild attempted and failed from a rebuild never attempted. Failed on attempt now demotes; not attempted blocks Tier 1 without demoting. Test 6 carries forward from t_0 to t_H, scored Failed on attempt at t_H where the rebuilder relied on a component that expires first.

Scope and definitions. The unit of assessment is clarified as the decision type and its record architecture rather than the firm, with a sampling note. Rebuild is distinguished from replay. The relationship to UK GDPR Article 5(1)(e) is stated.

Legal. The Article 86(1) qualifier, the full wording of Articles 19(1) and 26(6), the running point of Article 18(1) and the Digital Omnibus transitional rules have been added.

Editorial. References consolidated at section 9. Declarations added at section 8. UK punctuation conventions applied throughout.

Comment and criticism on this Assessment is invited. Cardinal AI Systems is the trading name of Whitehall Strategic Alliance Ltd.

Notes

1. Otopoetic, post beginning “Human observability is the condition that separates a governed AI system from a tolerated one”, LinkedIn, June 2026. https://www.linkedin.com/posts/otopoetic_otopoetic-ai-governance-classification-for-activity-7464935535283798016-6Ehg/ (accessed 15 August 2026). Archived copy: https://web.archive.org/web/20260815180358/https://www.linkedin.com/posts/otopoetic_otopoetic-ai-governance-classification-for-activity-7464935535283798016-6Ehg/ (captured 15 August 2026). Cited as one example of a formulation in circulation, not as authority and not offered as evidence of its prevalence. The

archived copy is given because a document arguing that expiring records cannot be relied upon should not rest its one sourced example on a page the author can withdraw.