reasoning-methodology-v1.0

ACCOUNTABILITY.AI

Methodology Standard

Reasoning Capture Methodology

Evidentiary Standards for AI Decision Records, Version 1.0

Document Type	**Methodology Standard**
Version	1.0 Founding Draft
Date	March 2026
Status	Public Review – Seeking Comment
Companion to	ADR Specification v0.1 (ISBN 978-1-7389042-0-4)
Canonical URL	accountability.ai/reasoning-methodology-v1.0
License	Creative Commons CC-BY 4.0
ISBN	978-1-7389042-1-1

> "The methodology that defines what counts as adequate AI reasoning must itself be adequate."

This document defines what constitutes adequate, compliant, and evidentiarily defensible reasoning capture for AI decision records. It is the methodology layer without which the ADR specification's technical infrastructure produces no evidentiary value. A hash-chained log of inadequate reasoning is still inadequate reasoning.

founding@accountability.ai **accountability.ai**

Table of Contents

---

01 Purpose and Scope

The ADR Specification v0.1 defines the fields that a compliant Agent Decision Record must contain. It defines cryptographic requirements that make those records tamper-evident. What it does not define – and what this document addresses – is what constitutes **adequate content** within those fields.

A reasoning field that is technically present but evidentially empty satisfies the letter of the specification while defeating its purpose. This methodology document defines the standards that distinguish adequate reasoning capture from inadequate reasoning capture, compliant records from non-compliant records, and evidence from mere documentation.

> **Core Principle**

> The technical infrastructure of the ADR – the hash chain, the Ed25519 signature, the canonical serialization – makes records tamper-evident. This methodology makes them evidentially meaningful. Both are necessary. Neither is sufficient alone.

1.1 What This Document Governs

This methodology applies to the following ADR fields:

- The `reasoning` field – the primary evidentiary field in any ADR
- The `reasoning_method` field – the method by which reasoning was generated
- The `feature_attribution` field – the quantitative basis for the decision
- The `confidence` field – the system's internal certainty metric
- The `input_summary` field – the evidence basis for the decision
- The `human_oversight` record – when human intervention occurs

1.2 What This Document Does Not Govern

This methodology does **not** govern:

- What AI systems may or may not decide – that is a matter for regulators, ethics bodies, and organizational policy
- What outcomes are acceptable – that is a matter for the jurisdictions listed in Section 09 of the ADR Specification
- How AI systems should be designed or trained – that is outside scope

The methodology governs only the evidentiary standard for recording decisions that have already been made.

---

02 The Evidentiary Standard

An AI decision record is evidentially adequate if and only if it satisfies **three criteria simultaneously**. A record that satisfies only one or two of the three criteria is not compliant with this methodology, regardless of technical conformance with the ADR Specification.

Element	Standard	Rationale
Contemporaneous	Captured at inference time	Post-hoc reconstruction of reasoning is not evidence. It is testimony. Testimony can be challenged on the grounds of memory, motivation, and reconstruction bias. A contemporaneous record cannot be so challenged if the cryptographic chain of custody is intact.

Element	Standard	Rationale
Specific	Tied to this decision, not decisions in general	Generic reasoning that would apply equally to any decision of this type is not specific. “Approved because applicant met criteria” is generic. It tells a reviewer nothing about which criteria, which values, and what the margin was.
Verifiable	Traceable to model outputs and inputs	Reasoning that cannot be traced to specific input features, confidence values, or model outputs cannot be independently verified. It is an assertion, not evidence.

2.1 The Contemporaneous Requirement in Detail

The reasoning field must be captured as part of the same transaction that produces the decision output. The timestamp of the reasoning must match the timestamp of the decision within the tolerance of a single inference call.

COMPLIANT	NON-COMPLIANT
Reasoning captured by the model at inference time as part of a chain-of-thought prompt, structured output, or explanation API, where the reasoning is written to the ADR record in the same atomic operation as the decision output.	Reasoning generated by a separate API call after the decision has been produced. Reasoning manually entered by a human reviewer. Reasoning generated by a different model than the one that made the decision. Reasoning templates populated with decision values after the fact.

2.2 The Specificity Requirement in Detail

Compliant reasoning must address **five specific elements**. A reasoning field that omits any of these elements is deficient, not merely incomplete.

1. **The decision itself** – what the system decided and at what confidence level
2. **The primary factors** – which input features drove the decision, in which direction, and with what weight
3. **The counterfactual threshold** – how far the inputs would need to change to produce a different decision
4. **Data quality** – any staleness, missingness, or quality issues present at decision time
5. **The policy basis** – which version of which policy authorized this decision

Elements 1, 2, and 5 are required for **all decisions**. Element 3 (counterfactual threshold) is required for decisions where confidence is below **80%**. Element 4 (data quality) is required when `data_quality_flags` is non-empty.

2.3 The Verifiability Requirement in Detail

Verifiable reasoning contains at least one **quantitative reference** that an independent reviewer can check against the model's outputs.

Examples of verifiable references:

- A confidence score: “Approved at 84% confidence”
- A feature weight: “Income-to-debt ratio contributed +23% toward approval”
- A threshold reference: “Score of 712 exceeds minimum threshold of 680”
- A comparison to policy: “Risk classification 'limited' per policy v2.3.1”

Reasoning that contains **no quantitative reference** cannot be independently verified and does not meet the verifiability standard.

---

03 Minimum Content Standards by Decision Type

The following standards define the minimum required content for the reasoning field for each primary decision type supported by the ADR specification. These are **floors, not ceilings**. Implementations are encouraged to exceed these standards.

3.1 Credit and Lending Decisions

`decision_type: credit_approval | credit_limit | rate_determination`

Required Element	Minimum Content Standard
Decision and confidence	State approved/denied/referred and the exact confidence percentage. Required for all credit decisions.
Primary factors	Name at least three input features that drove the decision, each with direction and weight.
Threshold disclosure	State the minimum score or ratio threshold applicable to this decision. Required when confidence is below 80%.
Adverse factors	For denied and referred decisions: list all factors that negatively affected the outcome, with quantitative impact.
Data quality	State the source and timestamp of all credit bureau data. Required when `data_quality_flags` is non-empty.
Policy reference	State the exact policy version (e.g., `credit-policy-v2.3.1`) that authorized this decision.

3.2 Export Control and Sanctions Screening

`decision_type: export_eligibility | sanctions_screen | jurisdiction_check`

Required Element	Minimum Content Standard
Match status	State whether a match was found, the match confidence, and the lists or databases screened. If no match: state which databases were screened and the screening timestamp.
Entity resolution	Where a potential match was identified, state the matching criteria, the fields that matched, and the confidence of the entity resolution decision.
Jurisdiction analysis	State all applicable jurisdiction restrictions and the basis for each restriction determination.
Screening parameters	State the fuzzy matching threshold used, the date of the watchlist used, and any jurisdiction-specific screening parameters applied.
Human review trigger	If human review is required: state the specific condition that triggered the requirement and the policy basis for that condition.

3.3 Clinical Triage and Healthcare Decisions

`decision_type: clinical_triage | care_pathway | diagnostic_support`

> **Regulatory Note**
>
> Clinical AI decisions are subject to FDA Software as a Medical Device (SaMD) guidance in the US and equivalent frameworks in other jurisdictions. The reasoning capture standard in this section is designed to satisfy the traceability requirements of those frameworks. **Human oversight is required for all clinical decisions; the `human_oversight` record is REQUIRED, not optional.**

Required Element	Minimum Content Standard
Clinical recommendation	State the recommendation, the clinical pathway triggered, and the urgency classification. If fast-track: state the specific clinical indicators that triggered fast-track classification.
Evidence basis	Identify the specific clinical parameters used, their values, and their reference ranges. Required: at least the primary vital or diagnostic indicator driving the recommendation.
Risk stratification	State the patient's risk tier, the factors that drove stratification, and the confidence level of the stratification.
Model limitations	State any patient characteristics that fall outside the model's validated population. Required when the patient's demographics or clinical profile differ materially from training data.
Human review requirement	State that human clinical review is required and identify the specific clinical or policy basis for that requirement. The

Required Element	Minimum Content Standard
	`human_oversight` record must reference this field.

3.4 Hiring and Employment Screening

`decision_type: hiring_screen | candidate_ranking | background_assessment`

Required Element	Minimum Content Standard
Screening outcome	State whether the candidate advances, is declined, or is flagged for human review. State the confidence level of the screening outcome.
Criteria applied	List all criteria applied in this screening, their relative weights, and the candidate's performance against each. **No criterion may be applied that is not listed here.**
Protected class abstention	Explicitly state that the following attributes were not used: race, gender, age, religion, national origin, disability status, or any proxy for these attributes. **This statement must appear in every hiring screen ADR.**
Comparative context	State where the candidate ranks relative to the screened pool on the primary criteria. Required when the screening outcome is decline.
Human review trigger	For any decision that may constitute an adverse employment action: state the specific trigger and the policy basis for requiring human review.

---

04 Non-Compliant Reasoning Patterns

The following patterns are explicitly **non-compliant** with this methodology. An ADR record exhibiting any of these patterns does not satisfy the evidentiary standard, regardless of technical conformance with the ADR Specification.

4.1 Generic Boilerplate

Reasoning that could apply equally to any decision of the same type, without reference to the specific inputs, values, or circumstances of this decision.

× Non-Compliant Example

"Application approved. Applicant met all required criteria. Decision made in accordance with current lending policy."

This reasoning contains no specific values, no feature references, no confidence score, and no policy version. It is indistinguishable from the reasoning for any other approval

decision and provides no evidentiary basis for review.

4.2 Post-Hoc Rationalization

Reasoning that is generated **after** the decision output, whether by the same model, a different model, or a human. Post-hoc reasoning is testimony, not evidence.

Detection Method

Post-hoc rationalization can be detected by comparing the timestamp of the reasoning generation to the timestamp of the decision output. If the reasoning was generated in a separate API call, or if the reasoning timestamp post-dates the decision timestamp by more than the latency of a single inference call, the reasoning is post-hoc and non-compliant.

4.3 Template Population

Reasoning generated by populating a fixed template with decision values. Template-populated reasoning fails the specificity standard because the template structure, not the model's reasoning process, determines what is captured.

× Non-Compliant Example

“Applicant [NAME] scored [SCORE] against threshold [THRESHOLD]. Decision: [OUTCOME].”

Even when populated with real values, template-generated reasoning fails to capture the model's actual reasoning process and is non-compliant.

4.4 Reasoning Without Attribution

Reasoning that states conclusions without identifying the inputs that drove those conclusions.

× Non-Compliant Example

“High-risk applicant. Multiple negative indicators. Recommendation: deny.”

What are the negative indicators? What makes them high-risk? What was the confidence level? Without answers to these questions, the reasoning cannot be independently verified and cannot serve as the basis for an adverse action notice.

05 Approved Reasoning Methods

The `reasoning_method` field must contain one of the following values. Each value corresponds to a method of reasoning generation that this methodology recognizes as capable of producing compliant reasoning, provided the minimum content standards of Section 03 are met.

Method	Standard	Rationale
`chain_of_thought`	Model generates step-by-step reasoning before producing the decision output, as part of a single inference call	Contemporaneous by construction. Preferred method for language model-based systems.
`shap`	SHAP (SHapley Additive exPlanations) values computed at inference time for the primary model output	Quantitative and verifiable. Preferred method for tabular and gradient boosted model systems.
`lime`	LIME (Local Interpretable Model-agnostic Explanations) computed at inference time	Acceptable when SHAP is computationally infeasible. Results must be stable across repeated calls.
`rule_trace`	Full trace of rules evaluated and their outcomes in a rule-based or hybrid system	Provides highest specificity for regulatory examination.
`attention`	Attention weight analysis from transformer-based models	Acceptable as supplementary evidence only. Not sufficient as a standalone reasoning method.
`integrated_gradients`	Integrated gradients attribution computed at inference time	Acceptable for deep learning systems where SHAP is not tractable. Must be computed at inference.

---

06 Reasoning Quality Assessment Framework

This section provides a structured framework for assessing the quality of reasoning captured in an ADR record. The framework is intended for use by internal compliance functions, external auditors, and professional attesters.

6.1 The Four-Dimension Assessment

Reasoning quality is assessed across four dimensions. Each dimension is scored as **Compliant, Deficient, or Non-Compliant**. A record that is Non-Compliant on any single

dimension does not meet the evidentiary standard.

Dimension	Assessment Standard
Completeness	Does the reasoning address all required elements for this decision type (Section 03)? Missing required elements = **Non-Compliant**. Missing recommended elements = **Deficient**.
Specificity	Does the reasoning contain specific values, features, and references tied to this decision? Generic language applicable to any decision = **Non-Compliant**.
Verifiability	Does the reasoning contain at least one quantitative reference that can be checked against the model's outputs? No quantitative reference = **Non-Compliant**.
Contemporaneity	Was the reasoning captured at inference time as part of the same transaction? Post-hoc reasoning = **Non-Compliant**.

6.2 Automated Quality Checks

The following checks should be implemented programmatically at ADR generation time. Records failing these checks should be flagged for human review before being accepted into the ledger. These checks are indicators, not absolute barriers – a record may fail a check yet still be compliant if it meets the substantive standards of Section 03.

Check	Threshold	Action	Rationale
Substantive length	Minimum 50 words. For high-risk decisions (clinical, credit, hiring), 80+ words is strongly recommended.	Flag for review	Extremely short reasoning (<50 words) is rarely adequate.
Numeric presence	At least one numeric value	Flag for review	Reasoning without any numbers cannot reference confidence scores, feature weights, or thresholds – all essential for verifiability.
Feature reference	At least two input features named	Flag for review	Generic reasoning (“applicant met criteria”) is non-compliant. Specific features must be identified.
Method valid	`reasoning_method` must be one of the approved values in Section 05	Reject immediately	Unrecognized methods cannot be validated.
Temporal alignment	Reasoning timestamp within 100ms of decision timestamp	Flag for review	Indicates possible post-hoc generation.

Check	Threshold	Action	Rationale
Confidence alignment	Confidence ≤ 100% and ≥ 0%. If < 80%, counterfactual threshold must be present.	Flag for review	Confidence out of range indicates system error. Low confidence without threshold lacks specificity.
Policy reference	`policy_version` must match format of a registered policy in the agent registry	Flag for review	Unverifiable policy references undermine accountability.

Organizations may adjust thresholds based on their risk appetite and review capacity.

---

07 Human Oversight Record Standards

When a human reviewer intervenes in an AI decision, the quality of the human oversight record is subject to the same evidentiary standards as the AI reasoning record. A rubber-stamped approval is not adequate human oversight. An override with no documented basis is not adequate human oversight.

7.1 Compliant Human Review

A human oversight record is compliant if it satisfies **all** of the following:

1. The `reviewer_id` is traceable to a named, credentialed individual in the access-controlled reviewer registry
2. The `reviewed_at` timestamp follows the assignment timestamp by a duration consistent with genuine review – not instantaneous approval
3. The `reviewer_notes` field contains a **substantive basis** for the reviewer's decision, not a single word or standard phrase
4. When `outcome_type` is `modified`, the `modification_details` field is populated with a structured diff of what changed and why
5. The `review_duration_secs` field is populated and falls within the expected range for this decision type

7.2 Rubber-Stamp Detection

The following patterns indicate **rubber-stamping** and should trigger escalation to the compliance function:

- Review duration **below the minimum threshold** for this decision type (see below)
- Identical `reviewer_notes` across multiple records from the same reviewer
- Approval rate above **99%** for a single reviewer over any 30-day period
- Review duration variance **below 10 seconds** across 10 or more consecutive records

Decision Type	Minimum Review Duration
Credit approval	120 seconds minimum
Export eligibility	180 seconds minimum
Clinical triage	300 seconds minimum
Hiring screen	240 seconds minimum
Sanctions screening	90 seconds minimum

08 Versioning and Amendment

This methodology document uses the same semantic versioning as the ADR Specification. Version 1.0 is a founding draft released for public comment alongside ADR Specification v0.1.

Amendments to the minimum content standards in Section 03 constitute **minor** version increments.

Amendments to the evidentiary standard in Section 02 or the approved reasoning methods in Section 05 constitute **major** version increments.

Organizations that implement the ADR specification must reference the specific version of this methodology in their system registrations and attestation documents. Regulatory examinations will assess compliance against the methodology version in effect at the time each record was produced.

8.1 How to Engage

This methodology is published for public comment alongside the ADR Specification. Comments are particularly sought from:

Regulatory practitioners who can identify gaps in the minimum content standards

Expert witnesses in AI liability cases who can assess the courtroom adequacy of these standards

AI system implementers who can identify compliance barriers

Legal practitioners specializing in evidence and discovery who can evaluate the admissibility framework

Submit comments to: **spec@accountability.ai**