Auditable Decision Intelligence: A Decision-Centered Definition of Explainable AI for High-Stakes Decisions

AiNOS Research

June 2026

Abstract

High-stakes AI decisions are not made explainable by rationales, saliency maps, chain-of-thought traces, or citations. What deployment demands is *answerability*: that an institution can answer for a decision when it is challenged. This paper defines Auditable Decision Intelligence (ADI) as that condition—a consequential AI decision is explainable only when its source-to-decision record is at once reconstructable, contestable, governable, and bounded. The contribution is not a synthesis of existing accountability frameworks but a claim they do not make: explainability is a property of a *single decision's record under contestation*, and no prior framework—procedural recordkeeping, reviewable automated decision-making, recourse and contestability, human-oversight regulation, or knowledge-limits standards—requires all four properties together over that one object. A system that satisfies each severally can still fail ADI. From this condition we derive a minimal Decision Audit Contract: the smallest record that makes one decision auditable. The contribution is conceptual—we define the target condition and the record it implies, and make no benchmark or architecture-performance claim.

1 Introduction

An AI system denies a consequential request. It returns a fluent paragraph citing an official-looking source and a confident rationale. Six weeks later the decision is challenged, and the institution must show exactly which source was used, whether it was current and in scope, what rule it triggered, who was authorized to approve the action, and how the affected party can contest it—and discovers that the explanation it produced answers none of these. This is the gap between an explanation that reads well and a decision an institution can answer for. High-stakes deployment therefore raises a stricter question than the existing explainable AI literature [1, 2, 3, 4, 5] was built to answer: *when is an AI-supported decision explainable enough for an institution to use it, defend it, correct it, and be held accountable for it?*

In high-stakes settings, explanation is not merely a communication artifact. A clinical recommendation, insurance compliance decision, credit decision, logistics exception, hiring decision, or legal-risk assessment can affect rights, safety, financial exposure, and institutional accountability. The system must therefore preserve more than an explanation artifact: it must show what was decided, what evidence was admissible or missing, what constraints and approvals governed the action, how much model freedom was allowed, and what record remains.

This paper proposes Auditable Decision Intelligence as a deployment-grade definition of explainability for high-stakes AI decisions. The core claim is simple: in consequential settings, explainability is not satisfied by showing something about the model or by producing something that

sounds like a reason. It is satisfied only when a qualified reviewer can reconstruct how admissible evidence, governing constraints, uncertainty, and institutional authority led to the final decision.

Contributions.

1. We distinguish model explanation, user-facing explanation, and decision auditability as the answerability condition deployment demands.
2. We define ADI as a decision-system property (Section 2) and show it is not the union of existing accountability frameworks.
3. We introduce a minimal Decision Audit Contract specifying what must be recorded for a high-stakes AI decision to be externally auditable.
4. We derive adequacy tests and failure modes that distinguish genuinely deployable explanations from plausible but non-auditable explanation artifacts.
5. This is a conceptual contribution: we define the target condition and the minimal record it implies; we do not build or evaluate a system, and make no benchmark or architecture-performance claim.

2 Auditable Decision Intelligence: Definition and Scope

2.1 Definition

Auditable Decision Intelligence is the deployment-grade form of explainability for high-stakes AI decisions. By *deployment-grade* we mean what consequential deployment *demands of* a decision, not a certificate that it is *cleared for* deployment (Section 2.3).

> A consequential AI decision is explainable when a qualified reviewer can externally reconstruct and contest its source-to-decision record: what was decided, what evidence was admissible, how source claims became decision-relevant facts, what governed the action, how much discretion the process was permitted and by what route it reached the outcome, what uncertainty remained, and what record stands for review. These terms name general properties of any governed decision process; the definition is agnostic to whether they are realized by a workflow engine, a retrieval pipeline, a rule system, or human procedure.

The definition is intentionally decision-centered. It does not ask whether a model is generally transparent, whether a generated rationale is fluent, or whether a citation appears near a claim. It asks whether this decision can be answered for under review. That shift is the paper's central move.

This definition has four core properties.

Reconstructable The decision can be replayed from source facts, constraints, route, verifier checks, and final record.

Contestable A reviewer can identify the exact fact, source, rule, constraint, judgment, or omission to challenge.

Governable The system controls decision freedom, routing, verification, and approval before the model acts.

Bounded Claims are limited to the covered domain, valid source substrate, known uncertainty, and tested decision conditions.

These properties are jointly required. Reconstructability without contestability produces a transcript that cannot be challenged. Contestability without boundedness invites disputes over facts and sources that the system was never entitled to use. Boundedness without governability produces correct-looking records after uncontrolled decisions. Governability without reconstructability produces policy compliance without explanation. Answerability begins only when all four are present.

2.2 Deriving the Four Conditions

The four properties are not meant as a preference list. They follow from the minimum accountability problem created by a consequential AI decision. A high-stakes institution must be able to say what was decided, why it was allowed, what evidence carried the decision, what uncertainty remained, who could challenge it, and whether the system acted within its authority. Each ADI property covers one part of that accountability problem. These requirements restate, in XAI terms, long-standing technological-due-process expectations for consequential automated decisions: that affected parties can challenge adverse outcomes and that regulators can test the system [39, 36].

Deployment requirement	**Required ADI property**	**Failure if missing**
The decision must be answerable after the fact.	Reconstructable	Reviewers cannot replay the evidence, constraints, route, and final record.
Affected parties and auditors must know what can be challenged.	Contestable	The system produces an explanation, but no specific fact, rule, or judgment can be disputed.
The system must act only with appropriate discretion and approval.	Governable	The model may choose a route, action, or freedom level the institution never authorized.
The decision must stay inside its evidential and jurisdictional limits.	Bounded	The system can use stale, irrelevant, incomplete, or out-of-scope information while still sounding plausible.

Table 1: Derivation of ADI's four required properties from high-stakes deployment requirements.

This derivation also clarifies the status of the definition. ADI is not a claim that every deployable AI system must expose all internal computation. It is a claim that every consequential AI decision must preserve the external conditions under which the decision can be answered for, challenged, governed, and bounded.

2.3 Adequacy of the Definition

ADI is adequate only for a specific scope: consequential institutional decisions where there is an accountable stakeholder, an affected party or audit interest, an admissible evidence boundary, and a meaningful possibility of review. Outside that scope, simpler forms of explanation may be sufficient.

Within that scope, the four conditions are necessary and non-redundant. Reconstructability supplies replay; contestability supplies challenge; governability supplies authorized discretion; boundedness supplies limits. Removing any one leaves a distinct failure mode—as the pairwise cases above show—that no surviving condition restores.

Boundedness presupposes an ex-ante admissibility policy for the decision context—which source types, recencies, and jurisdictions count as in scope—declared independently of the particular record, so the scope check runs against a standing rule rather than a judgment made after the outcome is known.

Governability and boundedness both impose constraints but on disjoint objects. Governability constrains the *action and discretion* of the system—may it act, with what freedom, via what route, under whose approval. Boundedness constrains the *epistemic basis*—whether the evidence is admissible, in scope, and within known uncertainty. The two are independently realizable: an unauthorized route over admissible evidence is a pure governance failure, while a fully authorized route over stale evidence is a pure boundedness failure. That establishes their orthogonality rather than asserting it.

The definition is also limited. Satisfying ADI is sufficient only for deployment-grade explainability, not for truth, fairness, legality, or social acceptability. A decision can be auditable and still wrong, unfair, or unlawful. ADI's contribution is to specify the minimum answerability condition that must exist before those further judgments can be made.

2.4 Why Exactly Four

The four properties are necessary, and sufficient against the leading candidate fifth properties, *for the answerability condition*—not for justice, pedagogy, or operational adequacy, which the definition deliberately excludes (Section 2.3). We do not claim completeness against every possible axis; we show that each leading candidate folds into an existing property.

Intelligibility to the affected party. One might propose comprehensibility as a fifth property. We treat it as a delivery requirement *on* contestability for the affected-party role: a record that a qualified affected-party reviewer cannot act on fails contestability, rather than escaping the definition. Intelligibility is thus a consequence of contestability correctly scoped to a role (Section 2.5), not a separate axis.

Timeliness. A contestation handle delivered after recourse is foreclosed is not a usable handle. Timeliness is a binding-time constraint on contestability and governability, not a fifth property.

Integrity. A record that can be silently altered cannot be faithfully replayed, so tamper-evidence is presupposed by reconstructability; verifiable, tamper-evident records are exactly the substrate prior accountability work specifies for ex-post review [36]. We make this explicit in the reconstructable gloss: replay assumes the record is the one produced at decision time.

Each candidate either folds into an existing property or is presupposed by one; none introduces a new accountability obligation left uncovered. That is the sense in which four suffices against the leading candidates; whether some further axis exists is left open.

2.5 Qualified Reviewers

The phrase "qualified reviewer" is role-specific.

- An affected party needs plain contestation handles: what fact or rule drove the outcome and what evidence could change it.

- A domain expert needs source references, rule application, missing evidence, and uncertainty state.
- A compliance reviewer needs jurisdiction, source validity, approval path, and record completeness.
- A system auditor needs route selection, verifier checks, model version, evidence assembly, and provenance.

ADI does not require that every reviewer see identical detail. It requires that the decision system preserve enough structured information for each role to perform its function.

Qualification must be fixed independently of the record under audit, or the definition is circular. A *qualified reviewer* for a decision and a role is one who possesses (i) the role's mandate and access rights and (ii) the domain and procedural competence standard fixed ex ante for that role—independent of the particular record. Examples: a licensed adjuster for compliance review; lay literacy plus recourse access for the affected party. Qualification is a property of the reviewer, not of their success on the record. A decision satisfies ADI when *any* reviewer meeting the fixed standard for a role, given only the preserved record, can reconstruct and contest it. This makes the condition falsifiable: failure by a qualified reviewer counts as an ADI failure rather than being explained away as the reviewer's incompetence, and a record an affected-party reviewer cannot act on fails contestability.

2.6 What Counts as an Explanation

Under ADI, an explanation is not the text emitted to the user. It is the availability of an auditable answer to three questions:

1. **Why was this decision allowed?** The answer must identify the decision objective, permitted action space, authority, rules, and approval requirements.
2. **Why was this evidence sufficient?** The answer must identify the admissible sources, source-derived claims, missing evidence, conflicts, and uncertainty boundaries.
3. **How can this decision be challenged or corrected?** The answer must expose the facts, source interpretations, constraints, judgments, and route choices that can be contested.

Natural-language summaries, saliency maps, counterfactuals, citations, tool traces, and model cards can help answer these questions. They are not by themselves the definition.

3 Existing Explanation Families and the Deployment Gap

The term "explainability" is overloaded. Classical interpretability work asks what makes a model or prediction understandable. Human-centered XAI asks what different users need from explanations. Evaluation work asks whether explanations are faithful, useful, stable, compact, or actionable. LLM research asks whether rationales or chain-of-thought traces reveal how language models arrive at answers. Retrieval-augmented generation asks whether generated outputs are grounded in sources. AI governance asks whether systems can be documented, audited, risk-managed, and overseen.

These literatures overlap, but they are not interchangeable. A model may be interpretable but used outside its intended context. A rationale may be plausible but unfaithful. A retrieved source may be cited but not decision- relevant. A governance document may describe a system but fail to

reconstruct a particular decision. ADI begins from the gap between these partial explanations and the stronger needs of deployed consequential decision making. Empirically, deployed explainability has been found to serve internal model debugging by ML engineers far more than the affected parties or auditors of consequential decisions [44], underscoring the gap between available explanation artifacts and the needs of decision review.

3.1 Classical XAI and Interpretability

Foundational work has shown that interpretability is purpose-dependent rather than a single universal property. Lipton highlights the ambiguity of interpretability. Doshi-Velez and Kim call for a rigorous science of interpretable machine learning. Arrieta et al. organize XAI concepts, taxonomies, and responsible AI challenges. Murdoch et al. distinguish predictive accuracy, descriptive accuracy, and relevance. Rudin argues that high-stakes decisions should prefer inherently interpretable models when possible rather than explaining unnecessary black boxes [2, 3, 1, 4, 5].

This literature supports one premise of ADI: high-stakes deployment requires more than post-hoc plausibility. But it does not define the full source-to- decision audit record required for deployed AI systems.

3.2 Local Explanation Artifacts

Local explanation methods such as LIME, SHAP, Anchors, and counterfactual explanations provide important artifacts for understanding individual predictions. They ask questions such as which features were locally influential, which rule-like conditions support a prediction, or what minimal change would alter an outcome [6, 7, 8, 9].

These methods are useful but not equivalent to deployment-grade explanation. Their unit of analysis is usually a prediction, feature contribution, saliency map, rule, or counterfactual. ADI's unit of analysis is a governed decision: what decision was requested, which source substrate was admissible, how source claims were transformed into decision-relevant facts, which constraints and approval obligations applied, which route was allowed, and what record remains. Local explanation artifacts may be included in an ADI record, but they cannot replace the record.

3.3 Human-Centered Explanation

Human-centered XAI emphasizes that explanation is audience-relative. Miller's social-science review and Mittelstadt et al.'s account of explanation show that explanations are selective, contrastive, and shaped by social and cognitive goals. Liao, Gruen, and Miller translate user needs into questions people ask about AI systems [10, 12, 11].

ADI inherits the audience-relative lesson but shifts the primary audience from a generic user to a set of audit roles: affected party, domain expert, compliance reviewer, and system auditor. A deployment-grade explanation must support these roles in reconstructing and contesting the decision.

Contestability is not only a user-interface preference. Work on algorithmic recourse, procedural fairness, and sociotechnical abstraction shows that affected parties need more than an explanation they can read; they need a handle through which a decision can be questioned, reversed, or corrected [14, 15, 16, 17]. Contestability has since been conceptualized as a system being open and responsive to dispute across its lifecycle [41, 40, 42]; ADI localizes this to per-decision contestation handles—the specific facts, sources, rules, judgments, and omissions a reviewer can challenge—treating them as part of the explanation object, not as an optional appeal feature.

3.4 Evaluation and Faithfulness

The XAI evaluation literature warns that explanation quality cannot be reduced to user satisfaction. Nauta et al. show that many XAI papers evaluate with anecdotal evidence and propose multiple conceptual properties for quantitative evaluation. NLP faithfulness surveys emphasize that an explanation should accurately reflect the basis of a prediction, not merely sound plausible [13, 18, 19].

ADI uses this distinction but applies it at the decision-system level. The question becomes not only whether an explanation is faithful to a model, but whether the decision record is faithful to the admissible evidence, governing constraints, routing policy, uncertainty state, verifier checks, and final action. Here faithfulness is to the *externally governed* decision process—admissible evidence, constraints, route, checks, and final action—not to the model's internal computation, which ADI does not assume is recoverable (Section 3.5).

3.5 LLM Rationales and Chain-of-Thought

Large language models can produce fluent rationales and chain-of-thought reasoning. These artifacts can improve performance, debug interaction, and help users inspect intermediate steps. But they are not stable audit records. Turpin et al. show that chain-of-thought explanations can systematically misrepresent the true reason for an answer. Faithfulness work in NLP and chain-of-thought reasoning shows why the stated reasoning path must be evaluated rather than assumed [18, 20, 21, 19].

ADI therefore does not rely on hidden chain-of-thought as the basis for deployable explanation. The audit object should be an externally inspectable record of admissible sources, claims, constraints, route, checks, and final decision.

3.6 RAG, Grounding, and Citation

Retrieval-augmented generation improves factual grounding by conditioning generation on external sources. RAG is important for ADI because high-stakes decisions should not rely on untraceable parametric memory. However, citation and retrieval alone do not establish deployment-grade explanation. A system can cite a source while misreading it, omit a required source, retrieve evidence that is relevant but not authoritative, or fail to explain how a source changes the decision.

Recent RAG and citation-evaluation work sharpens this point. RAG improves access to non-parametric memory, ALCE evaluates citation quality, RAGAS separates retrieval and generation quality, and attribution work shows that correctness and faithfulness are distinct properties [22, 23, 24, 25, 26]. ADI treats retrieval as necessary infrastructure, not as the complete explanation.

The relevant baseline is therefore not weak RAG. A strong RAG system may use retrieval logs, citation checking, abstention, reranking, structured extraction, and source attribution. ADI's claim is narrower: even strong RAG meets the answerability condition only when the decision it supports is itself auditable in the sense of Section 2—that is, when retrieval is one input to a governed decision rather than the whole of the explanation.

3.7 Accountability, Audit, and Governance

AI accountability literature shifts attention from model outputs to lifecycle processes, documentation, and institutional responsibility. Model Cards, Datasheets, Data Statements, FactSheets, provenance standards, algorithmic impact assessments, transparency critiques, and ethics-based auditing all reflect a common insight: consequential AI systems require structured evidence about

intended use, data, evaluation, risk, responsibility, and traceability [28, 29, 30, 31, 32, 27, 34, 35, 33, 36].

Standards and regulatory frameworks point in the same direction. NIST's Four Principles of Explainable AI emphasize explanation, meaningfulness, explanation accuracy, and knowledge limits. NIST AI RMF frames explainability as one trustworthiness characteristic among accountability, transparency, safety, privacy, fairness, reliability, and security. The EU AI Act and ISO standards stress risk management, transparency, human oversight, recordkeeping, and management-system governance [45, 46, 47, 48, 49, 50].

The common signal across these standards is not a single technical method. It is that high-stakes AI must preserve traceability, oversight, risk boundaries, recordkeeping, responsibilities, and knowledge limits. ADI translates that standards pressure back into XAI terms as the four-property condition of Section 2, applied to one decision.

This is why ADI is not merely another documentation artifact. Model Cards, Datasheets, and FactSheets describe systems, models, or datasets. Provenance standards describe derivation and attribution. Algorithmic audits examine organizational accountability. ADI complements these by specifying what must be available for one consequential decision to be reconstructed and contested.

Across these families the same gap recurs: an explanation can be plausible yet rest on the wrong basis, cited yet not decision-relevant, correct yet offer no contestation handle, interpretable yet routed wrongly, or source-grounded yet out of scope. Each is a failure of the decision record, not of the explanation's surface quality—these per-family insufficiencies are what the failure taxonomy (Section 5) makes scoreable.

3.8 Reviewability, Contestability, and Why ADI Is Not a Relabeling

The strongest objection to ADI is that each property renames an existing requirement. Closest is the *reviewability* framework of Cobbe et al. [37], which argues that accountable automated decision-making requires contextually appropriate records across the socio-technical process so a decision can later be reviewed. Kroll et al. [36] establish decision-level accountability through procedural regularity and verifiable records. Henin and Le Métayer [38] argue that explainability is insufficient for legitimacy and that justifiability and contestability are the binding requirements; the contestability-by-design line [42, 40, 43, 41] operationalizes contestation across the system lifecycle; and counterfactual recourse supplies a contestation handle [9]. At the standards level, EU AI Act Articles 12 and 14 [48] mandate lifetime logging and human oversight, and NIST's knowledge-limits principle [45] requires a system to declare its operating envelope.

ADI shares these targets but is not their union. The distinction is the *object*: prior frameworks govern the system or lifecycle (reviewability, EU Art. 12/14, NIST RMF) or the affected-party appeal (recourse, contestability-by-design). None binds a *single decision* into one object whose record must be at once reconstructable, contestable, governable, and bounded under contestation. Table 2 makes the residual concrete.

This yields a falsifiable claim: a system fully compliant with Kroll, Cobbe reviewability, EU AI Act Art. 12/14, and NIST IR 8312 can still fail ADI. Concretely, take one adverse eligibility decision. The system holds complete lifetime logs (EU Art. 12) under a standing oversight capability that approved the run (Art. 14); a procedurally regular, tamper-evident record exists (Kroll); reviewability records span the lifecycle (Cobbe); and the declared operating envelope was respected (NIST IR 8312). Yet the stored record sets the source document beside the adverse outcome with no recorded link from the source span to the claim to the eligibility boundary it crossed. Every prior requirement is met, but a qualified reviewer cannot contest *which* fact drove *this* outcome: the decision fails ADI's contestability over its own record. This decision-impact link is not entailed by a

ADI property	Closest prior requirement	What prior gives	Residual ADI adds
Reconstructable	Procedural regularity; reviewability records [36, 37]	System/lifecycle records, ex-post review	Per-decision source-to-claim-to-boundary replay
Contestable	Recourse; contestability-by-design [9, 38, 40]	A right or channel to challenge	Handles fixed to the same record's specific facts, rules, omissions
Governable	Human oversight [48]	Oversight is possible	Authorized freedom, route, and approval recorded per decision
Bounded	Knowledge limits [45]	Operating envelope declared	Source admissibility and scope checked for this decision

Table 2: The residual each ADI property adds over its closest prior requirement.

faithful reading of Cobbe's contextually appropriate records or Kroll procedural regularity, both of which are satisfied by the lifecycle record above, and it is distinct from attribution, groundedness, and provenance: an attribution can be correct without being faithful to what drove the decision [26], groundedness binds a claim to a source rather than to the boundary it crossed [25], and provenance records derivation rather than per-decision impact [32]. The Decision Audit Contract's contribution is therefore not its field list, every field of which exists in prior schemas, but the required co-presence of all four properties and the decision-impact link (Section 4) that provenance and documentation standards do not require.

4 The Decision Audit Contract

The Decision Audit Contract is the minimum structured object required for deployment-grade explanation. It is not a prompt template and not hidden reasoning. It is the audit object that remains after a decision.

The contract is intentionally minimal. A field belongs in the minimum contract only if removing it would prevent a qualified reviewer from reconstructing the decision, identifying a contestation target, determining whether the model had appropriate freedom, or verifying whether abstention, approval, or routing was required. Domain-specific systems may add richer logs, UX explanations, model internals, or organizational documentation, but those are extension layers.

All rows are minimum fields: each row's "failure if omitted" names a property broken that no surviving field recovers. Two pairs invite a merge objection. Source claims and decision-impact links stay distinct because a claim can be admissibly sourced yet have no recorded effect on the decision (a true but irrelevant fact, a reconstructable gap), while a decision can hinge on a boundary whose supporting claim is unrecorded (a contestable gap); neither field recovers the other. Verifier checks and approval requirement stay distinct because a semantic check and a human-authority gate fail in different ways and are owned by different roles. The criterion is thus non-self-referential: a field stays separate if and only if it has a failure mode no other field recovers, which the two witnesses above exhibit. A domain may legitimately merge or split rows only when this criterion permits.

Field	ADI property served	Failure if omitted
Decision objective	Reconstructable	Reviewers cannot tell what consequential decision was requested.
Decision type	Governable	The system may use the wrong freedom level or route.
Admissible sources	Bounded	Evidence may be stale, unauthorized, or outside scope.
Source claims	Reconstructable	Load-bearing facts cannot be checked against sources.
Decision-impact links	Contestable	Reviewers cannot see how a source changed the outcome.
Action rules and constraints	Governable	The action space and invalid actions are invisible.
Freedom level	Governable	The model may exercise unauthorized discretion.
Route	Reconstructable	Reviewers cannot tell whether code, rules, model reasoning, verifier, or human approval was used.
Missing/conflicting evidence	Bounded	The system can proceed despite gaps, conflicts, or uncertainty.
Verifier checks	Governable	Required deterministic or semantic checks can be bypassed.
Approval requirement	Governable	Mandatory human approval can disappear from the record.
Final decision	Reconstructable	The audit object has no stable outcome to review.
Contestation handles	Contestable	No fact, rule, constraint, judgment, or omission can be challenged.

Table 3: Minimal Decision Audit Contract for deployment-grade explainability.

4.1 A Decision Record Under Contestation

Consider a consequential eligibility, safety, or compliance decision. A fluent explanation might say that the request was approved because the applicant met the relevant criteria. Under ADI, that is not yet deployably explainable. The record must show which decision was requested, which sources were admissible, which source claims established each criterion, which constraints limited the action space, which missing evidence or conflicts remained, which verifier checks ran, whether approval was required, and which facts or judgments can be challenged.

If a reviewer contests one load-bearing claim, the record should expose the source span, the transformation from source to claim, and the link from claim to decision boundary. If that link is absent, the decision may still be correct, but it is not explainable under ADI. If a required source or approval step is missing, the explainable outcome may be escalation or abstention rather than a completed answer.

5 Failure Taxonomy

ADI makes failure modes scoreable.

Failure	Meaning
Unsupported claim	A load-bearing claim lacks admissible source support.
Source misread	The cited source is interpreted incorrectly.
Irrelevant source	The source is real but not decision-relevant.
Missing evidence ignored	Required information is absent but the system proceeds as if complete.
Conflict ignored	Conflicting sources or facts are not surfaced.
Wrong decision type	The system misclassifies the nature of the decision.
Wrong freedom level	The model receives too much or too little discretion.
Invalid action space	The selected output was not an allowed action.
Verifier bypass	Required checks were skipped or not recorded.
Approval bypass	Human approval was required but not routed.
Out-of-scope source	The source is stale, jurisdictionally wrong, or outside the covered domain.
Non-contestable record	The final record does not expose challengeable facts, rules, or judgments.

Table 4: Failure taxonomy for non-deployable explanations.

6 The Negative Case: Abstention

One boundary case is worth stating on its own, because it inverts a common intuition. A high-stakes system can be *more* explainable by declining to decide than by producing a fluent answer—*provided the abstention itself is recorded as an ADI decision*: the record exposes the missing evidence, the residual uncertainty, and the authority to which the case is escalated. A bare refusal is not explainable merely because it withholds an answer; it must satisfy the same four conditions. The other apparent boundary cases—a transparent model on an invalid source basis, a cited answer with a weak source relation, a fluent rationale with no contestation handle—are already resolved by the failure taxonomy (Section 5): each fails a specific ADI property despite looking explainable. A definition earns its place by changing how such cases are judged, and this one does.

7 Evaluation Implications

ADI turns explainability into an evaluable property of decision records. We run no evaluation here; the point is that the definition is operationalizable in principle. Because each Decision Audit Contract field is directly checkable, evaluation reduces to asking, per decision, whether each contract field is present, faithful, and contestable for the relevant review role—a target distinct from user satisfaction, prediction accuracy, citation presence, or rationale fluency, and one that broad transparency does not meet unless the evidence-to-decision relation is preserved.

8 Definition-Level Implications

The definition yields three implications for high-stakes AI. *First, explainability is relational*—a relation among a decision, its evidence substrate, governing constraints, decision authority, affected parties, and qualified reviewers that no single artifact satisfies in isolation. *Second, explainability is negative as well as positive*: it values correct refusal, escalation, and uncertainty disclosure, so an unsupported answer is not more explainable for being more complete (Section 6). *Third, explainability is prior to implementation*—the target condition is what must be true for a decision

to count as explainable before any substrate (rule engine, workflow, provenance database, retrieval system, or human review) can claim to deliver it.

9 Ethical Considerations, Adverse Impacts, and Artifacts

ADI is intended to reduce institutional opacity, but it can be misused. A decision audit contract can become compliance theater if organizations record formal fields without preserving real source sufficiency, contestability, or approval discipline. It can also create a false sense of legitimacy if a well-structured record is treated as proof of substantive justice—which, as Section 2.3 states, auditability never establishes. ADI should be evaluated as an auditability contract, not as such a guarantee.

The adverse-impact risk is highest when the contract is deployed to rationalize decisions that affected parties cannot practically challenge. A usable ADI implementation must therefore expose contestation handles, missing-evidence states, source validity limits, and approval obligations to the relevant review role, not merely store them in internal logs.

This paper does not collect human-subject data or introduce a deployed decision system. Its artifact is a conceptual definition, a decision audit contract, and a failure taxonomy. Any future empirical evaluation should report the domain, source substrate, review roles, scoring protocol, and abstention rules used to test whether decision records satisfy the ADI definition.

10 Limitations

ADI is not a universal theory of explanation. It does not explain model parameters, replace mechanistic interpretability, or claim that source-grounded decisions are always true. Sources can be stale, incomplete, conflicting, or jurisdictionally wrong. For that reason, boundedness, source validity, conflict handling, and correct abstention are part of the definition.

ADI should be evaluated domain by domain. Each domain must define admissible sources, action spaces, approval rules, auditor roles, and contestation procedures. The general contribution is the contract shape; the local implementation must be domain-specific.

Finally, ADI defines the contestation *artifact*—the record an affected party can challenge—but not the power, resources, or recourse access needed to use it; that access is assumed, not delivered. Absent it, the contract risks the compliance theater warned against above, where well-structured records coexist with decisions affected parties cannot practically contest.

11 Conclusion

High-stakes AI needs explanations that can survive audit, not only explanations that sound plausible. The central contribution of Auditable Decision Intelligence is to define deployment-grade XAI as a decision-system property: reconstructable, contestable, governable, and bounded source-to-decision answerability. This definition preserves the value of model interpretability, human-centered explanation, chain-of-thought research, RAG grounding, documentation, and AI governance while showing why none is sufficient alone. The takeaway is that high-stakes explainability is not the presence of an explanation artifact. It is the ability to answer for a decision under contestation.

References

[1] A. Barredo Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020. doi:10.1016/j.inffus.2019.12.012.

[2] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 2018. doi:10.1145/3233231.

[3] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.

[4] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *PNAS*, 2019. doi:10.1073/pnas.1900654116.

[5] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019. doi:10.1038/s42256-019-0048-x.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. *KDD*, 2016. doi:10.1145/2939672.2939778.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017. arXiv:1705.07874.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. *AAAI*, 2018. doi:10.1609/AAAI.V32I1.11491.

[9] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 2018. doi:10.2139/ssrn.3063289.

[10] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019. doi:10.1016/j.artint.2018.07.007.

[11] Q. V. Liao, D. Gruen, and S. Miller. Questioning the AI: Informing design practices for explainable AI user experiences. *CHI*, 2020. doi:10.1145/3313831.3376590.

[12] B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in AI. *FAT**, 2019. doi:10.1145/3287560.3287574.

[13] M. Nauta et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 2023. doi:10.1145/3583558.

[14] S. Venkatasubramanian and M. Alfano. The philosophical basis of algorithmic recourse. *FAccT*, 2020. doi:10.1145/3351095.3372876.

[15] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon. Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. *CHI*, 2023. doi:10.1145/3544548.3581161.

[16] A. D. Selbst and S. Barocas. The intuitive appeal of explainable machines. *Fordham Law Review*, 2018. doi:10.2139/ssrn.3126971.

[17] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. *FAT**, 2019. doi:10.1145/3287560.3287598.

[18] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *ACL*, 2020. doi:10.18653/v1/2020.acl-main.386.

[19] Q. Lyu, M. Apidianaki, and C. Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 2024. doi:10.1162/coli_a_00511.

[20] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388, 2023.

[21] T. Lanham et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702, 2023.

[22] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 2020. `https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html`.

[23] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. *EMNLP*, 2023. doi:10.18653/v1/2023.emnlp-main.398.

[24] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. *EACL*, 2024. doi:10.18653/v1/2024.eacl-demo.16.

[25] A. Stolfo et al. Groundedness in retrieval-augmented long-form generation: An empirical study. *Findings of NAACL*, 2024. doi:10.18653/v1/2024.findings-naacl.100.

[26] J. Wallat, M. Heuss, M. de Rijke, and A. Anand. Correctness is not faithfulness in RAG attributions. arXiv:2412.18004, 2024.

[27] I. D. Raji et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAccT*, 2020. doi:10.1145/3351095.3372873.

[28] M. Mitchell et al. Model cards for model reporting. *FAT**, 2019. doi:10.1145/3287560.3287596.

[29] T. Gebru et al. Datasheets for datasets. *Communications of the ACM*, 2021. doi:10.1145/3458723.

[30] E. M. Bender and B. Friedman. Data Statements for Natural Language Processing: Toward mitigating system bias and enabling better science. *TACL*, 2018. doi:10.1162/tacl_a_00041.

[31] M. Arnold et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 2019. doi:10.1147/JRD.2019.2942288.

[32] W3C. PROV-DM: The PROV Data Model. W3C Recommendation, 2013. `https://www.w3.org/TR/prov-dm/`.

[33] J. Mokander, J. Morley, M. Taddeo, and L. Floridi. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 2021. doi:10.1007/s11948-021-00319-4.

[34] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish. Algorithmic impact assessments and accountability. *FAccT*, 2021. doi:10.1145/3442188.3445935.

[35] M. Ananny and K. Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 2018. doi:10.1177/1461444816676645.

[36] J. A. Kroll et al. Accountable algorithms. *University of Pennsylvania Law Review*, 2017. `https://pennlawreview.com/2017/02/23/accountable-algorithms/`.

[37] J. Cobbe, M. S. A. Lee, and J. Singh. Reviewable automated decision-making: A framework for accountable algorithmic systems. *FAccT*, pp. 598–609, 2021. doi:10.1145/3442188.3445921.

[38] C. Henin and D. Le Métayer. Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI & Society*, 2022. doi:10.1007/s00146-021-01251-8.

[39] D. K. Citron. Technological due process. *Washington University Law Review*, 2008.

[40] K. Alfrink, I. Keller, G. Kortuem, and N. Doorn. Contestable AI by design: Towards a framework. *Minds and Machines*, 2023. doi:10.1007/s11023-022-09611-z.

[41] H. Lyons, E. Velloso, and T. Miller. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2021. doi:10.1145/3449180.

[42] M. Almada. Human intervention in automated decision-making: Toward the construction of contestable systems. *ICAIL*, 2019. doi:10.1145/3322640.3326699.

[43] K. Vaccaro, K. Karahalios, D. K. Mulligan, D. Kluttz, and T. Hirsch. Contestability in algorithmic systems. *CSCW Companion*, 2019. doi:10.1145/3311957.3359435.

[44] U. Bhatt et al. Explainable machine learning in deployment. *FAT**, 2020. doi:10.1145/3351095.3375624.

[45] P. J. Phillips et al. Four principles of explainable artificial intelligence. NISTIR 8312, 2021. doi:10.6028/NIST.IR.8312.

[46] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. doi:10.6028/NIST.AI.100-1.

[47] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, 2024. doi:10.6028/NIST.AI.600-1.

[48] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, 2024. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`.

[49] ISO/IEC. ISO/IEC 23894:2023 Artificial intelligence - Guidance on risk management, 2023. `https://www.iso.org/standard/77304.html`.

[50] ISO/IEC. ISO/IEC 42001:2023 Artificial intelligence - Management system, 2023. `https://www.iso.org/standard/42001`.