

Reliability of Audit Evidence in Artificial Intelligence Environments: Towards an Explainability and Traceability Evaluation Framework in Light of the 2026 Proposed Revisions to ISA 500

HADJAZE Khadidja¹, GUEMAZI Noudjoume²

¹ Setif 1 University Ferhat Abbas, (Algeria) , Email: khadjaze@univ-setif.dz ORCID: 0009-0005-5320-8598

² Associate Professor A, Setif 1 University Ferhat Abbas, (Algeria), Email: guemazibag@univ-setif.dz
ORCID:0009-0009-3261-1416

Received: 10/02/2026 | **Accepted:** 16/09/2026 | **Published:** 31/10/2026

ABSTRACT

Audit firms increasingly use machine-learning tools to analyze entire populations of transactions, yet their outputs are not in themselves audit evidence. This conceptual study examines how the reliability of AI-generated or AI-assisted information intended to be used as audit evidence can be evaluated systematically in light of the exposure draft of ISA 500 (Revised) issued by the IAASB in August 2026. The draft proposes that auditors evaluate the relevance and reliability of such information but leaves criteria for technology-produced information to judgment. The study combines a targeted critical review of auditing and explainable-AI research, a paragraph-level comparison of extant ISA 500 with the 2026 proposal, and design-oriented framework development. Its contribution is the Explainability–Traceability Audit Evidence Evaluation Framework (ET-AEEF), which organizes the evaluation around five linked dimensions: information provenance and integrity, algorithmic explainability, process traceability and reproducibility, evidential relevance and reliability, and professional judgment with human verification. It maps the proposal's reliability attributes onto observable criteria and applies a gate-and-compensation decision logic calibrated to the evidential role of the output. A hypothetical journal-entry anomaly-detection scenario shows that explainability alone cannot justify corroborative or primary reliance without traceability, and that an explainable, reproducible anomaly score directs further procedures rather than proving misstatement. The framework gives auditors and standard setters a transparent basis for documenting reliance decisions; it has not been empirically validated.

Keywords: Audit evidence; Automation bias; Explainable artificial intelligence ; ISA 500 ; Professional skepticism; Traceability

1. INTRODUCTION

Technology is changing not only how auditors gather evidence but also what they are asked to evaluate. When a machine-learning model scores every journal entry of a reporting period, the engagement team receives an object that did not exist in paper-based auditing: a ranked list of items whose selection logic the auditor did not observe and may not be able to reconstruct. Adoption is no longer the open question. Fedyk et al. (2022, p. 938) associate

How to cite: Khadidja & Noudjoume (2026). Reliability of Audit Evidence in Artificial Intelligence Environments: Towards an Explainability and Traceability Evaluation Framework in Light of the 2026 Proposed Revisions to ISA 500. *Reading in Political Economy*, 60(2), 1252–1262.

© 2026 The Author(s). Published by Reading in Political Economy.

audit-firm investment in artificial intelligence (AI) with a lower likelihood of restatements, and practitioners interviewed by Kokina et al. (2025, p. 1) describe document-extraction and similar tools as widely used while more complex applications remain under development. The open question is evidential: on what basis may the output of such a tool be relied upon when it shapes audit conclusions?

Algorithmic output strains evidential reliability in several distinct ways. The population processed by a tool may be incomplete or transformed before the auditor sees it. Complex models do not disclose why an item was flagged, and a model retrained between two runs may return different results for the same data. The behavioral risks run in both directions. Automation bias affects experts as well as novices and is not removed by training (Parasuraman & Manzey, 2010, p. 381), yet auditors in an experiment proposed smaller adjustments to management's estimates when contradictory evidence came from their firm's AI system rather than from a human specialist (Commerford et al., 2022, p. 171). Either deviation weakens professional skepticism.

The exposure draft of ISA 500 (Revised) issued by the International Auditing and Assurance Standards Board (IAASB) in August 2026, open for comment until 15 December 2026 (IAASB, 2026b, cover), makes these difficulties harder to ignore. It defines audit evidence as "information, after applying audit procedures, that the auditor uses to draw conclusions" (IAASB, 2026b, p. 27, para. 8(b)) and would require the auditor to evaluate the relevance and reliability of information intended to be used as audit evidence by reference to its source and to the attributes of reliability of significance in the circumstances (IAASB, 2026b, pp. 27–28, para. 11). At the same time, the Board considered and set aside a conditional requirement addressing the inputs, operation and outputs of technological tools, judging it premature to settle on a single workable model (IAASB, 2026b, p. 20, paras. 67–68). The proposal therefore poses a principles-based challenge without prescribing a technology-specific evaluation structure.

This article asks: *How can the reliability of AI-generated or AI-assisted information intended for use as audit evidence be systematically evaluated through explainability, traceability, and professional judgment in light of the 2026 proposed revisions to ISA 500?* Its contribution is the Explainability–Traceability Audit Evidence Evaluation Framework (ET-AEEF), a five-dimension evaluation structure with a purpose-calibrated decision logic. Explainability and traceability are treated throughout as the author's proposed evaluation dimensions, not as requirements of the proposed standard. The article moves from the literature (Section 2) and the standard (Section 3) to method (Section 4), framework (Section 5), illustration (Section 6), discussion (Section 7) and conclusion (Section 8).

2. LITERATURE REVIEW AND RESEARCH GAP

2.1 Audit evidence reliability and professional judgment

Extant ISA 500 defines audit evidence as information used by the auditor in arriving at the conclusions on which the opinion is based (IAASB, 2009, p. 3, para. 5(c)) and asks the auditor to consider the relevance and reliability of information to be used as audit evidence (IAASB, 2009, p. 4, para. 7). Its guidance on reliability takes the form of generalizations about source, nature and the circumstances in which evidence is obtained (IAASB, 2009, pp. 8–9, para. A31). These generalizations were written for confirmations, documents and inquiries, where the source is identifiable and the form of the evidence is fixed. The 2022 exposure draft moved toward attribute-based reasoning, suggesting characteristics such as accuracy, completeness, authenticity, bias and credibility (Barr-Pulliam et al., 2024, p. 38). Principles-based standards place the weight of the evaluation on the engagement team's judgment, which is where technology introduces new pressures.

2.2 Artificial intelligence in external-audit procedures

Evidence on the effect of AI on audit outcomes is encouraging but indirect. Fedyk et al. (2022, p. 938) link AI investment to fewer restatements and report, from partner interviews, that firms develop AI centrally and pursue it mainly for quality. Kokina et al. (2025, p. 1) find that the obstacles named by practitioners are transparency and explainability, bias, data privacy, robustness and reliability, over-reliance by auditors, and the absence of guidance. The review by Barr-Pulliam et al. (2024, pp. 46–47) adds that analytics can generate false positives, and it concludes that evidence gathered through technology is best suited to a complementary rather than primary role (Barr-Pulliam et al., 2024, p. 53). The literature thus disagrees less about whether AI helps than about how much evidential weight its outputs can bear, and it leaves the conditions for greater weight largely unspecified.

2.3 Explainable AI and algorithmic interpretability

Zhang et al. (2022, p. 1) introduced explainable AI (XAI) to auditing, demonstrating LIME and SHAP on a risk-of-material-misstatement task and arguing that such techniques can help satisfy documentation and audit evidence requirements. SHAP assigns each feature an importance value for a particular prediction (Lundberg & Lee, 2017, p. 4765), and the explanation is itself an interpretable approximation of the original model (Lundberg & Lee, 2017, p. 4766). The AI risk management framework of the National Institute of Standards and Technology (NIST) distinguishes explainability, which answers how a decision was made, from interpretability, which answers why and what it means in context (NIST, 2023, p. 17). Two critiques limit the evidential value of post hoc explanation. Slack et al. (2020, p. 180) show that perturbation-based methods such as LIME and SHAP can be fooled into presenting innocuous explanations for biased models, and Rudin (2019, p. 206) argues that high-stakes decisions should rely on inherently interpretable models rather than explained black boxes. Auditing research has not resolved what depth of explanation is sufficient for a given evidential purpose.

2.4 Data provenance, traceability, and reproducibility

A separate stream treats documentation as the basis of accountability. NIST (2023, p. 16) links the maintenance of training-data provenance to transparency and accountability; Raji et al. (2020, p. 33) propose an end-to-end internal audit of AI development in which each stage yields documents that together form an audit report; Mitchell et al. (2019, p. 220) propose model cards that disclose intended use and benchmarked evaluation; and Pineau et al. (2021, p. 1) define reproducibility as obtaining similar results with the same code and data. Li and Goel (2025, p. 1) derive an AI auditability framework from interviews with 23 AI professionals. These contributions principally concern AI governance or the audit of AI systems, rather than supplying a purpose-calibrated reliance procedure for outputs used in a financial-statement audit.

2.5 Automation bias, professional skepticism, and human oversight

Behavioral research identifies two opposite failures. Automation bias leads users to accept automated output with insufficient scrutiny, among experts as well as novices (Parasuraman & Manzey, 2010, p. 381), and the 2026 proposal names it explicitly (IAASB, 2026b, p. 35, para. A19). Algorithm aversion leads auditors to discount contradictory evidence produced by AI (Commerford et al., 2022, p. 171; Barr-Pulliam et al., 2024, p. 50). Skepticism therefore cannot be a general attitude toward technology; reliance should track evidence about the particular tool, run and purpose.

2.6 Critical synthesis and research gap

Three literatures run in parallel: auditing research on standards and adoption, XAI research on techniques and their fidelity, and AI-governance research on documentation for developers. Barr-Pulliam et al. (2024, p. 36) reviewed the 2022 draft, not the 2026 proposal with its revised definition and attribute-based evaluation requirement; Zhang et al. (2022, p. 1) demonstrate techniques but do not connect them to reliability attributes or to a decision procedure; Li and Goel (2025, p. 1) address the audit of AI systems rather than AI as a source of audit information. Within the purposively reviewed literature, a gap remains: the identified studies do not integrate the 2026 draft's reliability attributes with observable explainability and traceability criteria, differentiated evidential roles, and explicit reliance decisions. ET-AEEF addresses this defined gap without claiming an exhaustive absence of competing frameworks.

3. ISA 500 AND ITS 2026 PROPOSED REVISION

3.1 Status and architecture of the proposal

The 2026 Exposure Draft is a consultation document; its effective date is a placeholder (IAASB, 2026b, p. 27, para. 6), and the Board envisages possible approval in December 2027 with application to periods beginning at least 18 months after approval (IAASB, 2026a, p. 24, para. 79). Nothing in it is binding. Within the package, ED-500 supplies the reference framework for judgments about audit evidence, while proposed ISA 330 (Revised) and ISA 520 (Revised) address performance (IAASB, 2026a, p. 12, para. 30). It also builds on the 2022 draft: respondents then supported the input-output view of evidence but asked what constitutes a necessary audit procedure (IAASB, 2026b, p. 9, para. 13), and that draft had been deliberately non-prescriptive about technology (IAASB, 2026b, p. 20, para. 65).

3.2 From information to audit evidence

The central change is the input-output model. Information becomes audit evidence only after audit procedures are applied to it (IAASB, 2026b, p. 8, paras. 11–12; p. 29, para. A1). For AI-assisted work this distinction is decisive. A list of anomalous journal entries is, at most, information intended to be used as audit evidence; evidential status arises from the procedures the auditor applies to it and to the entries it identifies. The proposal also separates two levels of evaluation. Paragraph 11 concerns the relevance and reliability of particular information, whereas sufficiency and appropriateness, which the Board retained as well-embedded concepts (IAASB, 2026b, p. 9, para. 16), concern the body of evidence. The overall conclusion would move to ISA 700 (Revised) (IAASB, 2026b, pp. 9–10, para. 18), and proposed ISA 330 (Revised) adds a holistic stand-back evaluation (IAASB, 2026a, pp. 19–20, para. 54). An apparently reliable AI output, by itself, is never sufficient for an audit conclusion. Proposed paragraph 10 would require evaluation of whether the evidence obtained meets each intended purpose (IAASB, 2026b, p. 27, para. 10). Figure 1 represents this logic for AI-generated information, with the verification and skepticism checkpoints that precede any evidential use.

3.3 Reliability attributes and their significance

Application material lists accuracy, completeness, authenticity, absence of bias and credibility as attributes that may be considered (IAASB, 2026b, pp. 29–30, para. A4). The draft does not expect each attribute to be assessed individually (IAASB, 2026b, p. 14, para. 36); the auditor identifies those of significance in the circumstances, at least one, and the list is not intended as a checklist (IAASB, 2026b, pp. 15–16, paras. 41–42). For information produced by the entity, accuracy and completeness are presumed to be of significance (IAASB, 2026b, p. 16, para. 46; p. 28, para. 12). Evidence about reliability may come from identifying controls or from other procedures (IAASB, 2026b, p. 28, para. 13). Doubts trigger modified procedures, with an asymmetry: for relevance the auditor determines whether modification is possible, for reliability what modification is necessary (IAASB, 2026b, p. 28, para. 15; p. 19, para. 55). Inconsistent evidence must be addressed rather than set aside (IAASB, 2026b, p. 29, para. 17).

Figure 1. From AI-generated information to audit evidence: the evaluation logic of the 2026 proposed ISA 500

Information sources

General ledger, sub-ledgers, external data

AI processing

Technological tool: model, version, parameters

AI-generated output

Information intended to be used as audit evidence (para. 11)

Checkpoint 1

Source verification: authenticity, completeness

Checkpoint 2

Explainability and traceability of the run

Checkpoint 3

Skepticism: automation bias, algorithm aversion (A19)

Audit procedures

Relevance and reliability evaluated; doubts addressed (paras. 11–15)

Audit evidence

Information after applying audit procedures (para. 8(b))

Professional conclusion

Intended purpose met (para. 10); overall stand-back

No direct path: AI output becomes audit evidence only after audit procedures (para. A1); inconsistencies are investigated, not overridden (para. 17).

Note. Author's representation of the input-output model in IAASB (2026b, p. 8, paras. 11–12; p. 27, para. 8(b); p. 29, para. A1).

3.4 Technology, skepticism and documentation

Audit procedures may be performed manually or with technological tools (IAASB, 2026b, pp. 32–33, para. A14), and Appendix 3 offers non-mandatory considerations for their use (IAASB, 2026b, p. 33, para. A15; p. 21, paras. 70–71). The package would replace the phrase "automated tools and techniques" with "technological tools" (IAASB, 2026a, p. 21, para. 61). Procedures must not be biased toward corroborative evidence (IAASB, 2026b, p. 27, para. 9(a)). The Board chose not to add documentation requirements specific to ISA 500 (IAASB, 2026b, p. 17, paras. 49–50). Table 1 summarizes the comparison. The draft would require evaluation of AI-produced information intended for evidential use while leaving technology-specific criteria to judgment. Explainability and traceability are not named requirements, and ET-AEEF proposes them as a proposed way to operationalize paragraph 11 if finalized.

Table 1. Extant ISA 500 and the 2026 proposed revision: implications for AI-assisted procedures

Issue	Extant ISA 500	2026 Exposure Draft (proposed)	Implication for AI-assisted procedures
Definition of audit evidence	Information used by the auditor in arriving at conclusions (para. 5(c))	Information after applying audit procedures (para. 8(b); A1)	AI output is an input; evidential status arises only from procedures applied to it
Relevance and reliability	Auditor shall consider (para. 7)	Auditor shall evaluate, by reference to source and attributes of significance (para. 11)	Evaluation must address how tool inputs and processing affect the significant attributes
Reliability guidance	Generalizations on source, nature and circumstances (A31)	Accuracy, completeness, authenticity, absence of bias, credibility; not a checklist (A4; EM paras. 36, 41–42)	Attributes need translation into technical indicators of data and model quality
Evidence about reliability	Evaluate whether entity information is sufficiently reliable (para. 9)	Controls or other procedures (para. 13); accuracy and completeness to the extent necessary (para. 12)	Ingestion reconciliation and tool-validation evidence become relevant
Doubts and inconsistencies	Determine modifications or additions (para. 11)	Paras. 15–17; whether versus what modifications (EM para. 55)	Conflicts between AI flags and other evidence require investigation, not override
Technology	Information in electronic form (A12–A13, A31)	Tools may be used (A14); Appendix 3 guidance; no conditional tool requirement (EM paras. 67–71)	Explainability and traceability remain matters of judgment
Skepticism and bias	No reference to automation bias	Procedures not biased (para. 9(a)); automation bias named (A19)	Calls for checks against over- and under-reliance

Note. Paragraph references from IAASB (2009, pp. 3–9) and IAASB (2026b, pp. 8–35). EM = explanatory memorandum to the 2026 Exposure Draft. The proposed provisions are not effective.

4. RESEARCH METHODOLOGY

The study is conceptual and design-oriented and uses three complementary approaches. First, a targeted critical review draws on peer-reviewed auditing, explainable-AI and AI-governance publications, alongside original IAASB and NIST documents. Sources were selected for substantive relevance to evidential reliability, algorithmic explanations, traceability or professional reliance; search concepts included “audit evidence”, “artificial intelligence”, “explainable AI”, “automation bias”, and “data provenance”. The source base covers auditing standards and research published through October 2026, together with relevant foundational technical studies. This is a purposive synthesis, not a systematic or PRISMA review: reproducible database hit counts and screening logs are not claimed.

Second, a normative comparative analysis read extant ISA 500, the 2026 Exposure Draft with its explanatory memorandum, and the overall project memorandum at paragraph level, coding seven issues (Table 1). Third, the framework was developed in five steps: deriving evaluation obligations from the draft; mapping each reliability attribute to technical assurance properties identified in the literature; specifying indicators and procedures for each dimension; specifying a decision logic consistent with paragraphs 10, 15 and 17; and illustrating the result with one hypothetical scenario. Each construct is traceable to its source in Table 2. Validity is argued analytically, through traceability to authoritative sources and internal consistency. The framework has not been empirically validated, and all scenario facts are hypothetical.

5. THE PROPOSED ET-AEEF FRAMEWORK

5.1 Design principles

Three author-proposed principles govern the framework. The *evidential-status principle* holds that AI output remains information intended to be used as audit evidence until procedures have been applied to it, consistent with

the input-output model (IAASB, 2026b, p. 29, para. A1). The *purpose-calibration principle* holds that the required rigor depends on the evidential role the output is expected to play, consistent with the intended-purpose requirements (IAASB, 2026b, p. 27, paras. 9(b), 10). Three roles are distinguished: U1, informational use, in which output guides risk identification or the selection of items for testing; U2, corroborative use, in which output supports other evidence about an assertion; and U3, primary use, in which validated output may influence the extent of other procedures only after an overall assessment of evidence sufficiency. The *asymmetry principle* holds that unresolved failures in some dimensions preclude reliance on that output, while others may be addressed through documented independent procedures. These thresholds are author-proposed, not ISA mandates.

5.2 The five dimensions

D1. Information provenance and integrity is the degree to which the data processed by the tool can be traced to an authorized source and shown to be complete and unaltered through each transformation. It operationalizes authenticity and completeness (IAASB, 2026b, pp. 29–30, para. A4) and draws on the link between provenance and accountability (NIST, 2023, p. 16). Typical failures are partial extraction, silent filtering and unrecorded transformation. In ET-AEEF, an unverified input population blocks reliance on that specific output until independent reconciliation or re-extraction resolves the defect.

D2. Algorithmic explainability is the degree to which outputs relevant to the audit purpose can be interpreted and challenged by the engagement team. Global explanations describe what drives the model in general; local explanations, such as SHAP attributions for an individual entry (Lundberg & Lee, 2017, p. 4765), describe why a particular item was scored. An explanation describes the model, not the underlying transaction, and it may be unfaithful or manipulated (Slack et al., 2020, p. 180). D2 therefore informs scrutiny but never proves correctness.

D3. Process traceability and reproducibility is the ability to reconstruct the significant analytical steps, including data version, model version, parameters, thresholds, run time and user, and to assess repeatability or explain material variation across runs (Pineau et al., 2021, p. 1). Model cards and stage-by-stage documentation provide the necessary artifacts (Mitchell et al., 2019, p. 220; Raji et al., 2020, p. 33). Explanation without traceability is unverifiable, and traceability without explanation is reproducible opacity. In ET-AEEF, unresolved D3 failures bar U2 and U3 use.

D4. Evidential relevance and reliability translates technical findings into the terms of the standard. It asks whether the output bears on the relevant assertion and assessed risk, and which attributes are of significance (IAASB, 2026b, p. 28, para. 11(b)). Model validation evidence supports accuracy; bias in training data or thresholds bears on absence of bias; tool approval and the competence of its developers bear on credibility. Unresolved doubts lead to modified procedures (IAASB, 2026b, p. 28, para. 15).

D5. Professional judgment and human verification comprises independent corroboration, such as vouching flagged items and testing unflagged items for false negatives, the treatment of contradictory evidence (IAASB, 2026b, p. 29, para. 17), and explicit checks against automation bias and algorithm aversion. Accountability for the conclusion rests with the engagement team, not with the tool. Figure 2 shows how the dimensions interact: D1 conditions every later dimension; D2 and D3 jointly determine whether the output can be examined; D4 translates their results into reliability attributes; and D5 integrates them, either concluding or returning the evaluation to earlier dimensions through additional procedures. Table 2 gives the operational content of each dimension.

Figure 2. The proposed Explainability–Traceability Audit Evidence Evaluation Framework (ET-AEEF)

Intended purpose and evidential role (paras. 9(b), 10): U1 informational, U2 corroborative, U3 primary

D1 Information provenance and integrity

Source, completeness, unaltered transformations (gate)

D2 Algorithmic explainability

Global and local explanations; challengeable by the team (compensable for U1)

D3 Process traceability and reproducibility

Versions, parameters, logs; tested rerun (gate for U2, U3)

D4 Evidential relevance and reliability

Assertion and assessed risk; attributes of significance (para. 11(b)): accuracy, completeness, authenticity, absence of bias, credibility

D5 Professional judgment and human verification

Independent corroboration; contradictory evidence (para. 17); checks for automation bias and algorithm aversion; accountability of the engagement team

Decision

Further procedures, conditional use or rejection

Feedback loop: additional procedures, explanations or escalation (para. 15)

Documentation of ratings, rationale and limitations (R6)

Note. Author's proposal. Paragraph references are to IAASB (2026b); the framework is not prescribed by the IAASB.

Table 2. Operational dimensions, indicators, verification procedures and limitations of ET-AEEF

Dimension and attribute link	Observable criteria	Suggested verification procedures	Main failure risks	Decision implication
D1 Provenance and integrity (authenticity, completeness)	Record counts and totals reconcile to the ledger; source system and extraction documented; integrity hash	Independent extraction or reconciliation; review of transformation scripts	Partial or filtered population; altered fields	Gate: if inadequate, reject reliance
D2 Explainability (accuracy, absence of bias)	Global drivers documented; local attributions for flagged items; explanations intelligible to the team	Challenge explanations against business rationale; specialist review	Unfaithful or manipulated explanations; false comfort	U1: independent checks; U2: resolve D2 and corroborate; U3: adequate D2
D3 Traceability and reproducibility (credibility)	Versioned model and data; parameters, seed, threshold and user logged; rerun tested; differences explained	Re-performance or variance analysis; inspect logs and model card	Silent retraining; undocumented thresholds	Unresolved D3 bars U2/U3; qualify U1
D4 Relevance and reliability (attributes of significance)	Link to assertion and assessed risk; significant attributes identified; validation evidence	Map output to assertion; evaluate attributes; para. 13 procedures	Output relevant to a different risk; unexamined doubts	Doubts trigger para. 15 modification
D5 Judgment and human verification	Corroboration results; treatment of contradictions; bias checks	Vouching flagged items; testing unflagged items; consultation	Automation bias; algorithm aversion	Conclude, loop back, escalate or reject; document

Note. Author's operationalization. Attribute links refer to IAASB (2026b, pp. 29–30, para. A4). U1 = informational, U2 = corroborative, U3 = primary use.

5.3 Decision logic

Each dimension is rated adequate, partially adequate or inadequate. Qualitative ratings avoid unvalidated numerical precision; the Board likewise declined to define a fixed persuasiveness threshold (IAASB, 2026b, p. 10, paras. 19–20). The following conservative reliance rules are author-proposed, not ISA mandates. (R1) An unresolved D1 failure bars reliance on that output for the intended role; the source population must be independently verified or re-extracted. (R2) An unresolved D3 failure bars U2 and U3; any U1 use must disclose its limitations. (R3) Partial D2 may be tolerated for U1 with independent testing; U2 additionally requires an adequate further explanation or specialist challenge and corroboration, otherwise remains pending; U3 requires adequate D2. (R4) D4 doubts require modified procedures. (R5) D5 contradictions require investigation, never deference to the tool. (R6) Ratings, grounds and limitations must be documented. U3 further requires satisfactory evidence in the audit as a whole; passing these dimensions never automatically reduces substantive procedures. Figure 3 summarizes these decisions.

Figure 3. Risk-based decision pathway for evaluating AI-assisted audit information

Initial AI output

Anomaly scores and flagged items

1. SOURCE VERIFICATION (D1)

Population authentic, complete and reconciled?

2. EXPLAINABILITY ASSESSMENT (D2)

Outputs interpretable and challengeable for the purpose?

3. TRACEABILITY ASSESSMENT (D3)

Run reconstructable and reproducible?

4. RELEVANCE AND RELIABILITY (D4)

Bears on the assertion; significant attributes supported?

5. HUMAN CORROBORATION (D5)

Independent testing consistent with the output?

6. DECISION AND DOCUMENTATION

Evidential role fixed; ratings and rationale recorded (R6)

Yes

Yes

Yes

Yes

Yes

Reject reliance

Reperform on a verified population; output is non-evidential

No

Further procedures

Request explanation or escalate; then re-enter

Partly

Reject reliance for U2 and U3

Informational use only, with the limitation documented

No

Further procedures

Modify procedures (para. 15) or investigate inconsistencies (para. 17); never defer to the model; then re-enter the pathway

Doubts

Inconsistent

Conditional use

U1 directs testing; U2 needs R3 resolution; U3 needs adequate evidence

Note. Author's proposal. The pathway distinguishes further procedures, conditional use and rejection of reliance; it is not prescribed by the IAASB.

6. ILLUSTRATIVE APPLICATION AND ANALYTICAL FINDINGS**6.1 Hypothetical audit context**

All facts in this section are hypothetical. An engagement team auditing a mid-sized manufacturing company has assessed a risk of management override and plans to test journal entries. The ledger contains about 412,000 entries for the year. A gradient-boosted anomaly model scores each entry, and flags 1,240 entries at an illustrative, unvalidated threshold of 0.85. The team intends to use the flags to select entries for testing, which is informational use (U1), and is considering whether independent tests could substantiate any corroborative use (U2); the unflagged rate alone cannot do so. In the terms of Figure 1, the flags are information intended to be used as audit evidence, and they reach evidential status only through further procedures.

6.2 Two contrasting situations

In Situation A the tool is run by a third-party provider on a file extracted by the client's IT staff. Entry counts were not reconciled to the general ledger, no model version or parameter log exists, and a second run returns 1,187 flags instead of 1,240. Scores are delivered without explanations. D1, D2 and D3 are all inadequate. Under D4 the team cannot evaluate completeness or credibility, and the main risk under D5 is that the scores are accepted as objective because they are quantitative. Rules R1 and R2 apply: reliance is rejected. The team extracts and reconciles the population itself, selects entries using documented risk criteria, and escalates the tool issue to the firm's technology function.

In Situation B the model runs in the firm's approved environment. Entry counts and totals reconcile to the ledger, the input file is hashed, and a model card and run log record validation, limitations, version and parameters; a controlled rerun reproduces the flags. Local attributions highlight postings outside the finance function, weekend postings and unusual account pairings. The team checks a risk-selected sample of flagged entries against approvals and documents, examines selected unflagged entries for potential missed risks, and investigates a conflicting explanation. These steps illustrate verification, not an estimated false-negative rate. D1 and D3 are adequate, and D4 and D5 support the limited purpose of selecting items; D2 remains partial beyond U1 because feature attributions do not establish transaction truth. Under R3 the output may direct testing (conditional U1), but U2 remains pending a further explanation or specialist challenge plus corroboration, neither assumed complete here. No reduction of substantive testing (U3) is justified. Table 3 contrasts the situations.

Table 3. Application of ET-AEEF to the hypothetical journal-entry scenario

Dimension	Situation A: opaque output, weak traceability	Situation B: explainable, traceable, corroborated output
D1 Provenance	Inadequate: client extract not reconciled	Adequate: counts and totals reconciled; file hashed
D2 Explainability	Inadequate: scores only	Adequate for U1; partial for U2/U3
D3 Traceability	Inadequate: no logs; rerun differs	Adequate: versioned run; rerun identical
D4 Relevance and reliability	Cannot evaluate completeness or credibility	Relevant to override risk; significant attributes supported
D5 Human verification	Risk of automation bias	Risk-selected entries checked; conflict investigated
Decision	Reject reliance (R1, R2); reperform selection; escalate	Conditional U1 only; U2 pending R3; no U3

Note. All facts and values are hypothetical and illustrate the framework only.

6.3 Analytical findings

Three findings emerge from the comparison. First, Situation B is more suitable for further evidential evaluation not because it is explained but because it can be verified: reconciliation fixes the population, run documentation permits re-performance, and independent checks test the underlying transactions. Had Situation A supplied attributions without logs, the explanations would have referred to a run that could not be revisited. Second, the evidential role governs the required rigor. The same output can guide testing (U1) but cannot yet provide corroboration (U2) or primary support (U3). Third, in Situation B reliance moves from the model to the auditor's own procedures. An anomaly score, however well explained, remains a reason to examine an entry, not evidence that it is misstated.

7. DISCUSSION, IMPLICATIONS AND LIMITATIONS

ET-AEEF extends prior work in three ways. First, it converts the XAI techniques demonstrated by Zhang et al. (2022, p. 1) into evaluation criteria tied to the reliability attributes of the 2026 draft, and it adds the warning, drawn from Slack et al. (2020, p. 180) and Rudin (2019, p. 206), that explanations are evidence about a model rather than about transactions. Second, it gives operational form to the conclusion of Barr-Pulliam et al. (2024, p. 53) that technology-gathered evidence is best used as complementary evidence, by specifying the further verification needed before stronger uses can be considered. Third, it shifts the focus of the auditability literature (Li & Goel, 2025, p. 1; Raji et al., 2020, p. 33) from auditing AI systems to evaluating AI output as audit information.

For external auditors the framework provides a vocabulary for documenting reliance decisions. For audit firms it implies that tool approval should yield engagement-usable artifacts, such as reconciliation routines, model cards and run logs, in line with the proposed conforming amendments on technological resources (IAASB, 2026a, p. 21, paras. 61–65). For professional education it suggests training auditors to read explanations critically and to recognize both automation bias and algorithm aversion. For standard setters, the Board's judgment that a single workable model is premature (IAASB, 2026b, p. 20, para. 68) leaves room for non-authoritative guidance; illustrations of traceability evidence in Appendix 3 or in implementation material could be considered during the consultation. Implementation faces real obstacles: cost and specialist expertise, restricted access to proprietary algorithms, cybersecurity of data pipelines, and tools such as generative AI that may not fit the input-output model the draft assumes (IAASB, 2026b, p. 20, paras. 67–68).

The study has clear limits. The framework is conceptual and has not been empirically validated; the rating categories and rules are analytical proposals, not IAASB requirements or tested measures. A single hypothetical scenario cannot establish effectiveness across tools, entities or jurisdictions. The analysis rests on an exposure draft that may change before approval. Future research should test the framework through experiments on auditor reliance, field studies of tool documentation in firms, and inspection-based analysis once revised standards are effective.

8. CONCLUSION AND RECOMMENDATIONS

In answer to the research question, the study proposes a five-dimension evaluation, covering provenance, explainability, traceability, evidential relevance and reliability, and human verification, governed by a purpose-calibrated decision logic. The principal analytical findings are that AI output is information rather than evidence until procedures are applied; that unresolved provenance failures preclude reliance, traceability failures bar U2–U3, and partial explainability may be addressed for U1 only through independent checks; and that even an explainable, reproducible anomaly score supports the direction of further procedures rather than a conclusion of misstatement.

The theoretical contribution is the mapping of the draft's reliability attributes onto observable technical criteria and the gate-and-compensation logic linking them to decisions. In practice, engagement teams should state the intended evidential role of each AI output before evaluating it, reconcile the processed population independently, require documented, testable run records, and examine risk-selected unflagged as well as flagged items. Because the framework is conceptual and the standard is still a proposal, its rules are hypotheses; empirical validation and revision against the final text of ISA 500 (Revised) are the research priorities.

REFERENCES

- Barr-Pulliam, D., Calvin, C. G., Eulerich, M., & Maghakyan, A. (2024). Audit evidence, technology, and judgement: A review of the literature in response to ED-500. *Journal of International Financial Management & Accounting*, 35(1), 36–67. <https://doi.org/10.1111/jifm.12192>
- Commerford, B. P., Dennis, S. A., Joe, J. R., & Ulla, J. W. (2022). Man versus machine: Complex estimates and auditor reliance on artificial intelligence. *Journal of Accounting Research*, 60(1), 171–201. <https://doi.org/10.1111/1475-679X.12407>
- Fedyk, A., Hodson, J., Khimich, N., & Fedyk, T. (2022). Is artificial intelligence improving the audit process? *Review of Accounting Studies*, 27(3), 938–985. <https://doi.org/10.1007/s11142-022-09697-x>
- International Auditing and Assurance Standards Board. (2009). *International Standard on Auditing 500: Audit evidence* [Edition consulted: reproduction with IFAC permission by the Malaysian Institute of Accountants, January 2020]. International Federation of Accountants. <https://circulars.mia.org.my/2020/mf09/ISA%20500.pdf>
- International Auditing and Assurance Standards Board. (2026a). *Exposure draft for the Audit Evidence and Risk Response project overall: Explanatory memorandum and proposed conforming and consequential amendments*. International Federation of Accountants. <https://ifacweb.blob.core.windows.net/publicfiles/2026-08/IAASB-Audit-Evidence-Risk-Response-Overall-Project-Exposure-Draft.pdf>

- International Auditing and Assurance Standards Board. (2026b). *Exposure draft: Proposed International Standard on Auditing 500 (Revised), Audit evidence, with explanatory memorandum*. International Federation of Accountants. <https://ifacweb.blob.core.windows.net/publicfiles/2026-08/IAASB-Audit-Evidence-Risk-Response-ISA-500-Exposure-Draft.pdf>
- Kokina, J., Blanchette, S., Davenport, T. H., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. *International Journal of Accounting Information Systems*, 56, Article 100734. <https://doi.org/10.1016/j.accinf.2025.100734>
- Li, Y., & Goel, S. (2025). Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. *International Journal of Accounting Information Systems*, 56, Article 100739. <https://doi.org/10.1016/j.accinf.2025.100739>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pp. 220–229. Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164), 1–20. <http://jmlr.org/papers/v22/20-303.html>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pp. 33–44. Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (pp. 180–186). Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375830>
- Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable artificial intelligence (XAI) in auditing. *International Journal of Accounting Information Systems*, 46, Article 100572. <https://doi.org/10.1016/j.accinf.2022.100572>