

REVIEW

Open access

Measuring What Explanations Actually Achieve—An Umbrella Review of Fidelity, Stability, Comprehensibility, Decision Usefulness, Human Trust, and Reliance

Sarah Bennett¹, Kwame Asante^{2*}, Elena Rossi³

Abstract

Explainable artificial intelligence has generated extensive technical and human-centered evaluation research, yet substantial uncertainty remains about what explanation evaluations actually demonstrate. This umbrella review synthesizes review-level evidence concerning the evaluation of fidelity, faithfulness, stability, robustness, comprehensibility, cognitive accessibility, decision usefulness, human trust, trust calibration, and reliance. Rather than reviewing explanation algorithms as methods, the article examines the constructs, metrics, study designs, and validity assumptions through which explanation quality has been assessed. The synthesis distinguishes technical correspondence with model behavior from human judgments of plausibility, understanding, usefulness, and satisfaction. It further separates human trust from system trustworthiness and behavioral reliance from beneficial decision support. Review-level evidence indicates that technical metrics are heterogeneous, terminology is inconsistently operationalized, and stable or plausible explanations are not necessarily faithful. Human-centered evaluations frequently rely on self-reported understanding, satisfaction, or trust, while objective comprehension, cognitive load, decision quality, appropriate reliance, overreliance, and underreliance receive less consistent assessment. Evidence linking technical explanation properties to downstream human and institutional outcomes remains limited, particularly across stakeholder groups, realistic decision contexts, repeated use, and changing system conditions. The umbrella synthesis therefore rejects single-metric interpretations of explanation quality and organizes evaluation as a multidimensional evidence problem involving technical validity, cognitive accessibility, decision consequences, calibrated trust, reliance behavior, accountability, and contestability. Its principal contribution is a consolidated evaluation agenda that clarifies what existing measures may support, what they do not establish, and which methodological connections require stronger user-centered, task-specific, longitudinal, and socio-technical validation.

Keywords Explainable AI, Stability, Fidelity, Umbrella review, Explanation evaluation, Comprehensibility

*Correspondence:

Kwame Asante

kwame.asante@ug.edu.gh

¹ Department of Explainable AI and Human Factors, University of Manchester, Manchester, United Kingdom

² Department of AI and Decision Support, University of Ghana, Accra, Ghana

³ Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Introduction

Explainable artificial intelligence has developed from a collection of methods for exposing aspects of opaque model behavior into a broader research domain concerned with transparency, interpretability, human interaction, and responsible system use. Foundational surveys nevertheless reveal substantial heterogeneity in explanation objectives, method

families, terminology, and evaluation assumptions [1–3]. Consequently, the presence of an explanation cannot itself demonstrate that the explanation accurately represents a model, helps an intended user, improves a decision, or supports accountable system operation.

Explanation is also a cognitive and social phenomenon rather than solely a technical artifact. People often seek selective, contrastive, context-sensitive accounts that answer particular questions and fit existing knowledge, expectations, and practical goals [4]. An explanation that is mathematically detailed may therefore be inaccessible to its intended audience, whereas an intuitively plausible account may conceal important model behavior. Technical correspondence and human intelligibility must accordingly be treated as related but distinct evaluation targets.

The central evaluation problem is not whether an explanation can be produced, but what the explanation actually achieves. Relevant outcomes include fidelity to the predictive system, stability under appropriate perturbations, comprehensibility for intended stakeholders, support for accurate mental models, usefulness within a decision task, calibrated rather than maximized trust, appropriate rather than indiscriminate reliance, and effective opportunities for scrutiny or challenge. These outcomes occur at different evidentiary levels and should not be collapsed into a general claim that an explanation is “good.”

This article provides an original umbrella review of review-level evidence on explainability evaluation. It compares how prior reviews characterize technical properties, human-centered outcomes, trust, reliance, decision support, and evaluation validity; identifies recurrent construct and measurement gaps; and develops a future evaluation agenda. The synthesis does not independently establish that explanations improve understanding, decisions, accountability, or deployment outcomes. Its contribution is to clarify what existing evaluations measure, which conclusions those measures may support, and where stronger evidence remains necessary.

Evaluation Challenges in Explainable AI

Explainability evaluation is complicated because explanation systems are designed for different users and purposes. Developers may seek debugging information, auditors may require evidence of model behavior, domain professionals may need decision-relevant reasoning, and affected individuals may need understandable grounds for challenge or recourse. Evaluation should therefore be aligned with the intended stakeholder, explanatory objective, model context, and practical task rather than applying the same success criterion universally [5].

Terminological inconsistency further weakens comparison. Explainability, interpretability, transparency, comprehensibility, fidelity, faithfulness, robustness, usefulness, and trust are sometimes treated as interchangeable despite referring to different properties or outcomes. Reviews of evaluation approaches show that technical, human-grounded, and application-grounded assessments are often combined without clear construct boundaries [6]. Metric validity becomes uncertain when a measure's label, operational definition, and intended inference do not correspond.

Technical evaluation usually examines whether an explanation reflects some aspect of model behavior. Proposed measures address correctness, fidelity, robustness, consistency, completeness, sparsity, sensitivity, or related properties, but their definitions and implementations vary across explanation families. A systematic review of quantitative evaluation methods found broad metric proliferation alongside uneven coverage and limited standardization [7]. A favorable score may therefore support a narrow property claim while remaining insufficient for claims about comprehension, usefulness, trustworthiness, or decision quality.

Human-centered evaluation introduces additional validity problems. Self-reported clarity, satisfaction, confidence, or trust can be informative, but they do not necessarily demonstrate objective understanding, calibrated reliance, or improved decisions. Ecological validity, external validity, stakeholder diversity, system uncertainty, and repeated-use effects are also frequently underdeveloped. Explainability should consequently be treated as one component of trustworthy artificial intelligence rather than as independent proof of fairness, safety, accountability, or readiness for operational use [8].

Umbrella Review Scope and Review-Level Evidence Sources

The umbrella review addresses six questions: which evaluation challenges recur across review-level XAI evidence; how fidelity, faithfulness, stability, robustness, and consistency are assessed; how comprehensibility, cognitive accessibility, mental models, and understanding are evaluated; how decision usefulness, trust calibration, reliance, and automation bias are examined; which cross-dimensional patterns emerge; and which evidence gaps should guide future evaluation. A meta-survey perspective is appropriate because the objective is to synthesize patterns across reviews rather than reinterpret primary findings as direct umbrella-level proof [9].

Eligible evidence sources comprise peer-reviewed review articles and review-oriented syntheses that contribute directly to explanation evaluation. Systematic reviews, survey reviews, meta-surveys, comprehensive reviews, multidisciplinary syntheses, and selected state-of-the-art assessments were prioritized. Supporting conceptual or methodological studies were reserved for constructs requiring additional clarification. Reviews of user studies are especially important because they reveal variation in study populations, explanation tasks, outcome measures, experimental designs, and interpretations of trust or understanding [10].

Evidence extraction was organized around publication type, explainability context, technical evaluation properties, human-centered constructs, decision and trust outcomes, evaluation methods, principal limitations, and contribution to the umbrella synthesis. Broad XAI reviews were used to contextualize method and application diversity, but their findings were not treated as direct evidence that explanations produce beneficial human outcomes [11]. Categories not clearly reported by a source were treated as unreported rather than inferred.

Synthesis was qualitative and construct-centered. Review-level patterns were compared across fidelity, stability, comprehension, cognitive accessibility, decision usefulness, trust, reliance, accountability, and validity. Particular attention was given to tensions between technical and human outcomes and to emerging evaluation challenges involving foundation models. Contemporary interdisciplinary agendas support movement toward stakeholder-aware, impact-oriented, and longitudinal assessment while cautioning against further metric proliferation without stronger validity evidence [12].

Table 1 summarises the umbrella review logic used to identify, screen, map, and synthesize review-level evidence on explainable AI evaluation.

Table 1. Umbrella Review Logic for Explainable AI Evaluation: Review Questions, Evidence-Source Types, Inclusion Criteria, Exclusion Criteria, Screening Logic, Review-Level Evidence Verification, Evidence Extraction Categories, Evaluation-Dimension Mapping, Citation Planning, and Review Boundaries

Review component	Core review question or purpose	Inclusion logic	Exclusion logic	Evidence mapped	Evaluation dimension addressed	Bias or limitation risk	Umbrella review
Review questions	Determine what XAI evaluations measure and establish	Questions address technical, cognitive, behavioral, decision, and accountability outcomes	Questions limited to ranking explanation algorithms	Recurring review findings and limitations	Cross-cutting	Scope may become excessively broad	Structure the synthesis

Evidence-source types	Prioritize review-level knowledge	Systematic, survey, comprehensive, multidisciplinary, and meta-review sources	Non-peer-reviewed or method-only material without evaluation relevance	Review aims, methods, findings, and gaps	Review-level evidence	Review types use different methods	Support umbrella-level comparisons
Inclusion criteria	Retain sources supporting planned claims	Relevant peer-reviewed journal source with DOI and verified journal standing	Source lacks evaluation relevance or usable evidence	Technical or human-centered evaluation contribution	All domains	Selection may favor prominent terminology	Preserve claim-level relevance
Exclusion criteria	Prevent unsupported evidence inflation	Not applicable	Conference-only, preprint-only, legal, commercial, policy, or documentation sources	Publication status and substantive fit	Evidence validity	Relevant interdisciplinary evidence may be omitted	Protect source consistency
Screening logic	Assess relevance and eligibility without fabricated counts	Title, abstract, full-text, journal, DOI, and publication-type checks	Unverifiable or weakly fitting sources	Eligibility characteristics	Cross-cutting	Screening judgments may be subjective	Produce an audit trail logic
Review-level verification	Separate reviews from supporting studies	Review synthesis is clearly identified	Primary evidence represented as umbrella-level evidence	Review status and synthesis scope	Evidence interpretation	Reviews may reproduce weaknesses in primary studies	Maintain article-level integrity
Evidence extraction	Capture comparable constructs and limitations	Reported technical, human, decision, trust, reliance, or accountability evidence	Unreported information inferred by the reviewer	Aim, context, construct, method, limitation	Evaluation validity	Reporting detail varies	Enable structural comparisons
Dimension mapping	Link evidence to defined evaluation domains	Source supports at least one mapped construct	Decorative or random citation placement	Fidelity, stability, comprehension, usefulness, trust, reliance	Technical and human-centered	Constructs may overlap	Organize cross-review synthesis

Citation planning	Preserve first-appearance order and claim fit	Every source has a planned sentence and section	Unplanned grouped or weakly fitted citation	Claim-to-source relationship	Evidence traceability	Overcitation may obscure interpretation	Protected Vancouver sequence
Review boundaries	Prevent unsupported conclusions	Cautious synthesis and proposed future directions	Universal metrics, causal proof, certification, or readiness claims	Non-sufficiency boundaries	Validity and deployment boundaries	Readers may interpret synthesis prescriptively	Defined defensibility conclusions

Figure 1 illustrates the umbrella review evidence-mapping workflow used to move from review questions and review-level evidence screening to evaluation-dimension mapping, synthesis, and future evaluation agenda development.

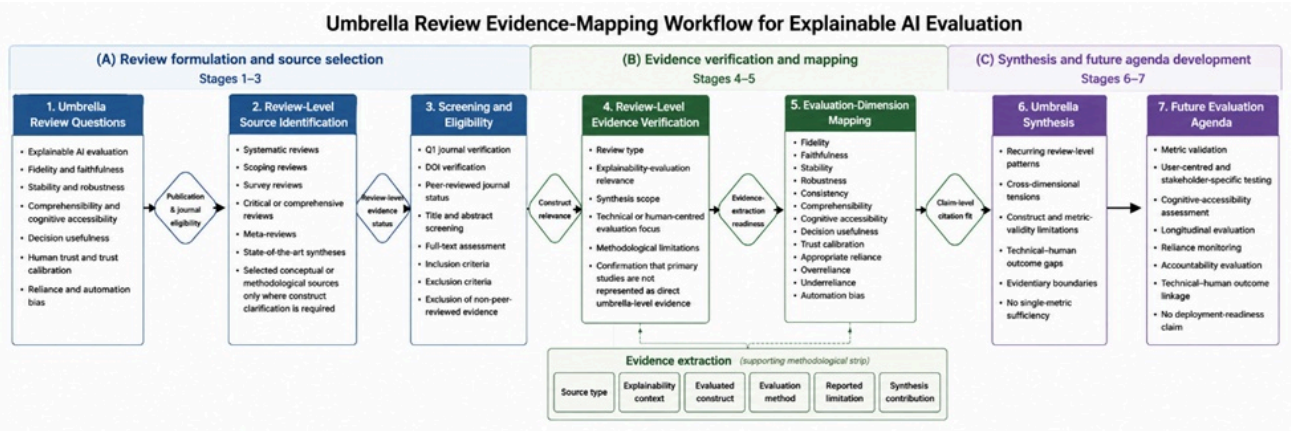


Figure 1. Umbrella Review Evidence-Mapping Workflow for Explainable AI Evaluation

Alt-text description:

An umbrella review workflow diagram showing review questions, review-level source screening, eligibility checks, evidence verification, evaluation-dimension mapping, evidence extraction, umbrella synthesis, and future evaluation agenda development

Fidelity, Faithfulness, and Stability Evidence

Technical evaluation asks whether an explanation corresponds to the behavior of the model it purports to explain. Reviews distinguish attribution, surrogate, perturbation, representation, and example-based assessments, while comparative benchmarks show that explanation methods can receive different rankings under different metrics or tasks [13, 14]. Fidelity scores therefore require interpretation relative to the explanation type, model scope, sampled region, perturbation design, and property being measured.

Fidelity commonly denotes agreement between an explanation or surrogate and model outputs, whereas faithfulness concerns whether the explanation reflects factors actually used by the model. The distinction becomes especially important for post-hoc explanations in high-stakes settings, where a plausible account may rationalize rather than reveal model behavior. Methodological critiques consequently warn against treating post-hoc explanation as equivalent to direct interpretability or transparent reasoning [15].

Model fidelity, explanation fidelity, causal faithfulness, completeness, and sensitivity address related but non-identical questions. A local surrogate can closely reproduce predictions within a restricted neighborhood without characterizing the wider model. Likewise, a complete mathematical decomposition may remain cognitively unusable. Technical surveys of neural-network interpretability show that validity criteria change according to whether the target is a prediction, feature attribution, internal representation, learned concept, or broader model mechanism [16].

Stability, robustness, and consistency concern explanation behavior under repeated computation, input perturbation, model retraining, or adversarial manipulation. Stability is desirable when changes are irrelevant, but sensitivity is necessary when changes are consequential to the model. Evaluation must therefore define which variations should preserve an explanation and which should alter it. In safety-critical contexts, technically polished explanations may otherwise create false reassurance without establishing causal validity, clinical usefulness, or reliable model behavior [17].

No technical property is sufficient by itself. Stable explanations may be consistently wrong; faithful explanations may be too complex for intended users; plausible explanations may be unfaithful; and simple explanations may be incomplete. Feature-importance agreement also provides weak evidence when methods share assumptions or when no independent ground truth exists. Technical evaluation should therefore use property-specific evidence while avoiding transitions from a narrow metric result to claims about human understanding, decision usefulness, trustworthiness, accountability, or deployment readiness.

Table 2 maps review-level evidence on fidelity, faithfulness, stability, robustness, consistency, completeness, plausibility, and technical evaluation limits in explainable AI.

Table 2. Review-Level Evidence on Fidelity, Faithfulness, Stability, and Technical Explanation Evaluation: Model Fidelity, Explanation Fidelity, Causal Faithfulness, Robustness, Consistency, Sensitivity, Completeness, Plausibility, Perturbation Sensitivity, Metric Validity, and Technical-Sufficiency Boundaries

Evaluation construct	Core technical evaluation question	Explanation property assessed	Potential evaluation approach	Review-level evidence pattern	Failure or overclaiming risk	Evidence gap
Fidelity	Does the explanation reproduce relevant model behavior?	Predictive or functional agreement	Surrogate agreement, deletion, insertion, or perturbation tests	Commonly evaluated but operationalized differently	Local agreement may be generalized beyond its tested scope	Weak cross-method comparability
Faithfulness	Does the explanation reflect factors actually used by the model?	Model-grounded explanatory accuracy	Ablation, intervention, randomization, or known-ground-truth tests	Often conflated with fidelity or plausibility	Persuasive rationalization may appear faithful	Limited independent ground truth
Causal faithfulness	Does the stated explanatory cause affect the outcome through the	Causal validity	Structural causal modelling and interventions	Rarely demonstrated directly	Correlation or perturbation may be mislabelled causal	Restrictive or absent causal assumptions

	claimed mechanism?					
Model fidelity	Does a surrogate approximate the source model in the relevant region?	Surrogate-to-model correspondence	Local or global prediction agreement and coverage	Highly dependent on sampled scope	Narrow fidelity may be presented as global fidelity	Inconsistent coverage reporting
Explanation fidelity	Does an explanation accurately characterize a particular prediction?	Instance-level correspondence	Feature removal, masking, counterfactual, or attribution tests	Metrics are explanation-family dependent	A metric may favor the assumptions of its own method	Limited cross-family validation
Stability and consistency	Are explanations reproducible across similar inputs, runs, or equivalent models?	Reproducibility and agreement	Rank correlation, repeated fitting, seed variation, or overlap	Definitions and neighbourhoods vary	Repeatability may be mistaken for correctness	No standard semantic neighbourhood
Robustness	Does the explanation resist irrelevant noise or manipulation?	Resistance to perturbation and attack	Noise, adversarial, retraining, and implementation stress tests	Post-hoc methods can be fragile	Robust appearance may conceal systematic error	Limited standardized stress testing
Sensitivity	Does the explanation change when model-relevant information changes?	Responsiveness to consequential variation	Controlled perturbation and ablation curves	Necessary complement to stability	Excessive stability may signal insensitivity	Few joint stability–sensitivity protocols
Completeness	How much relevant model behavior is represented?	Explanatory coverage	Reconstruction, attribution-sum, or coverage measures	Usually method-specific	Formal completeness may be cognitively unusable	Weak connection to user-relevant completeness
Plausibility	Does the explanation appear reasonable to a person?	Human-perceived coherence	Expert judgment or agreement with prior knowledge	Frequently treated as a quality proxy	Plausibility may conceal unfaithfulness	Few joint plausibility–faithfulness tests
Perturbation sensitivity	How does explanation output react to	Local response behavior	Masking, insertion, deletion, or	Widely used but design-sensitive	Unrealistic perturbations	Limited ecological validity

	controlled changes?		input-noise testing		can distort conclusions	
Metric validity	Does a measure capture the construct attributed to it?	Construct validity	Convergent, discriminant, predictive, and ground-truth validation	Validation remains uneven	Metric labels may imply unsupported properties	Few independently validated measures
No fidelity-sufficiency boundary	Does technical correspondence establish usefulness or understanding?	Boundary between technical and human validity	Combined technical and user/task evaluation	Review evidence does not support sufficiency	Faithful output may be unusable or harmful	Limited technical–human linkage studies
No stability-sufficiency boundary	Does repeatability establish correctness or value?	Boundary between stability and validity	Joint stability, sensitivity, fidelity, and outcome assessment	Stable explanations may remain wrong or irrelevant	Reproducibility may be treated as truth	Limited multidimensional assessment
No technical-metric sufficiency boundary	Can one score establish explanation quality?	Multidimensional adequacy	Purpose-specific metric portfolios	Reviews reveal persistent metric trade-offs	Universal scoring suppresses context and stakeholder differences	No validated universal composite metric

Comprehensibility and Cognitive Accessibility Evidence

Comprehensibility concerns whether an intended user can interpret an explanation accurately enough for its stated purpose. Medical-XAI reviews illustrate that explanation format, terminology, granularity, and visual representation must correspond to domain knowledge and clinical tasks rather than satisfy a generic simplicity criterion [18]. A representation that assists a model developer may be unsuitable for a clinician, patient, auditor, or person affected by an automated decision.

Understandability is often evaluated through perceived clarity, satisfaction, preference, or ease-of-use measures. These outcomes describe user experience but do not necessarily establish objective comprehension or accurate mental models. Healthcare evidence shows that terminology, design choices, and evaluation strategies remain inconsistent, limiting comparison between explanations and preventing broad inferences about trustworthy outcomes [19]. Evaluations should distinguish what users believe they understand from what they can correctly infer about system behavior, limitations, and errors.

Cognitive accessibility extends beyond nominal understandability. It includes the mental effort, time, literacy, expertise, accessibility requirements, and workflow conditions under which an explanation is processed. Multidisciplinary evidence indicates that explanation requirements vary across technical developers, healthcare professionals, organizational decision makers, and affected individuals [20]. Explanation evaluation should therefore consider whether users can access, interpret, and use the information under realistic constraints rather than only under idealized study conditions.

Stakeholder goals also shape the appropriate balance between simplicity and completeness. Lay users may require concise contrastive accounts, while technical experts may need detailed evidence for debugging and auditing. Domain

experts may require explanations that expose uncertainty, alternative hypotheses, and clinically consequential features. Stakeholder-oriented synthesis consequently shows that explanation quality cannot be judged independently of the explanatory need being served [21]. Preferred explanations may still be inaccurate, insufficient, or ineffective for the target task.

Cognitive load, explanation overload, and explanation mismatch remain comparatively underexamined. Human-centered design work suggests that explanations should fit professional workflow, information needs, interaction patterns, and available decision time [22]. Progressive disclosure, interactive questioning, and adaptable detail may improve accessibility, but these strategies require validation. Comprehensibility alone does not establish fidelity, usefulness, appropriate reliance, or trustworthy system operation, and exposure to an explanation does not guarantee an accurate or durable understanding.

Decision Usefulness, Trust, and Reliance Evidence

Human trust is a multidimensional attitude influenced by system performance, task characteristics, institutional context, user experience, and presentation. Reviews also reveal substantial variation in the definitions and measurements applied to trust in AI-enabled systems [23, 24]. Trust must therefore be distinguished from trustworthiness: the former concerns a person's attitude, while the latter concerns whether the system merits confidence based on evidence such as competence, robustness, accountability, and responsible governance.

Explanation effects on trust are heterogeneous. Clinical reviews report that explanations may increase, decrease, or leave trust unchanged depending on explanation type, user expertise, system performance, and study design [25]. A directional change in trust does not establish improvement because reduced trust may be appropriate when a system is unreliable, while increased trust may constitute overtrust when an explanation is persuasive but incomplete or misleading.

Decision usefulness requires evaluation beyond user attitudes. Review-level synthesis suggests tracing a sequence from explanation properties to user perceptions, cognitive responses, behavioral choices, and downstream consequences [26]. Relevant outcomes include error detection, task performance, decision quality, appropriate acceptance of correct advice, rejection of incorrect advice, and effective use of uncertainty. Satisfaction or perceived usefulness should not substitute for evidence that explanations improve the decision process.

Trust calibration concerns whether confidence varies appropriately with system capability and case-level correctness. Evidence from clinical decision-support research indicates that explanation class, presentation, workflow fit, and professional context can affect calibration differently [27]. Calibration should therefore be assessed through users' responses to both accurate and inaccurate AI recommendations rather than through average trust ratings alone.

Reliance is behavioral and can diverge from stated trust. Explanations may alter confidence without improving performance, or change acceptance behavior without helping users distinguish correct from incorrect recommendations [28]. Appropriate reliance requires correct acceptance and correct rejection; overreliance includes following erroneous advice, whereas underreliance includes rejecting valid assistance. Evaluation must also consider automation bias, uncertainty communication, verification behavior, and human–AI complementarity. No increase in trust or reliance guarantees better decisions.

Table 3 synthesises review-level evidence on comprehensibility, cognitive accessibility, decision usefulness, human trust, trust calibration, reliance, overreliance, underreliance, and automation bias.

Table 3. Review-Level Evidence on Human-Centered Explainability Evaluation: Comprehensibility, Understandability, Cognitive Accessibility, Cognitive Load, Mental Models, Decision Usefulness, Decision Quality, Human Trust, Trust Calibration, Appropriate Reliance, Overreliance, Underreliance, Automation Bias, and Reliance Boundaries

Human-centered evaluation construct	Core human or decision question	User, decision, or task context	Evaluation function	Review-level evidence pattern	Failure or misinterpretation risk	Overview validation need
Comprehensibility	Can the intended user interpret the explanation?	Expertise- and task-dependent	Perceived and objective comprehension tests	Commonly assessed, but definitions vary	Clarity may be mistaken for understanding	Separate report demonstrates comprehension
Cognitive accessibility	Can the explanation be processed under real constraints?	Time, literacy, disability, stress, and workflow	Accessibility and workload assessment	Less studied than general usability	Technically clear information may remain inaccessible	Inclusive content sensitivity testing
Cognitive load	How much mental effort does processing require?	Complex or time-critical tasks	Workload, response-time, or secondary-task measures	Increasingly recognized but inconsistently measured	Excess detail can impair verification	Link workload to decision outcomes
Mental-model support	Does the user learn system capabilities and limits?	Repeated human–AI interaction	Prediction, transfer, and error-anticipation tasks	Frequently claimed but rarely tested over time	Explanations may create inaccurate mental models	Longitudinal and transfer assessments
User expertise	How does knowledge shape explanation use?	Lay, technical, and domain users	Stratified outcome analysis	Expertise affects needs and interpretation	Aggregate findings can conceal subgroup failure	Report specific user groups
Decision usefulness	Does the explanation aid the decision process?	Goal-defined decision task	Error detection, information use, and task outcomes	Less often validated than perceived usefulness	A liked explanation may be practically ineffective	Compare meaningful task outcomes
Decision quality	Are final decisions more appropriate?	Human-only, AI-only, and team settings	Accuracy, error cost, or domain-defined quality	Results are context-dependent	Compliance may be mislabelled as improvement	Evaluate correct/incorrect advice
Human trust	What attitude does the user hold toward AI?	System-, user-, and context-dependent	Trust scales and qualitative accounts	Concepts and measures are heterogeneous	Higher trust may be unwarranted	Pair trust with system performance evidence
Trust calibration	Does confidence correspond to capability?	Variable system correctness	Case-level alignment and error discrimination	Increasingly emphasized but inconsistently operationalized	Mean trust does not show calibration	Include case studies and individual cases
Appropriate reliance	Does behavior match when AI should be used?	Collaborative decision support	Correct acceptance and rejection	Often conflated with trust	Greater use may reduce decision quality	Measure behavior

						independ of attit
Overreliance	Are incorrect recommendations accepted?	Plausible AI errors and time pressure	Agreement with wrong advice and missed-error rates	Explanations do not consistently prevent it	Persuasive rationales can strengthen deference	Include v advice
Underreliance	Are correct recommendations rejected?	Low trust or confusing explanations	Rejection of correct advice	Less studied than overreliance	Useful support may be unnecessarily ignored	Evalu disuse missed l
Automation bias	Does AI distort information search or judgment?	Authority cues and pressured decisions	Omission, commission, and verification measures	Effects vary with explanation design	Explanation may become another authority cue	Bias-se experim condit
No trust-sufficiency boundary	Does increased trust prove explanation quality?	All human–AI contexts	Joint trust and performance assessment	Evidence does not support the inference	Trust maximization can encourage overreliance	Asse whether warran
No reliance-benefit boundary	Does greater reliance improve outcomes?	Depends on AI correctness and complementarity	Conditional reliance analysis	Reliance may help or harm	Adoption may be mistaken for success	Link bel to dec qual

Umbrella Synthesis of Explainability Evaluation

Across the reviewed evidence, explanation evaluation forms a chain rather than a single measurement problem. Technical validity may be necessary for some purposes, but downstream outcomes depend on whether users engage with and verify explanations. Experimental evidence indicates that explanations may reduce overreliance when their information is usable and when users have sufficient motivation and ability to evaluate it [29]. Explanation presence alone is therefore a weak intervention specification.

Automation bias further complicates the relationship between explanation and decision usefulness. Review evidence suggests that explanations can support critical examination, but can also increase deference when they appear authoritative, fluent, or confirmatory [30]. Evaluation should distinguish assistance from cognitive substitution and should examine whether users search for independent evidence, recognize AI errors, and adjust reliance appropriately.

Trustworthy evaluation also extends beyond the explanation interface. Robustness, fairness, privacy, human oversight, security, and accountability remain relevant even when an explanation is technically faithful and understandable [31]. Explanation evaluation should consequently be embedded within broader system assurance rather than used as a substitute for evaluating the intelligent system, organizational process, or decision environment.

Accountability-oriented explanations require more than descriptive transparency. Recourse research distinguishes explanations of why an outcome occurred from recommendations that are feasible, actionable, and capable of producing a consequential change [32]. Contestability additionally requires that explanations support the identification of possible error, meaningful challenge, review by an appropriate authority, and correction or remedy where warranted.

Counterfactual proximity, sparsity, and plausibility may improve presentation but do not guarantee causal validity or feasible recourse [33]. Foundation-model explanations add further concerns because fluent generated rationales may be evaluated as explanations despite uncertain correspondence with internal mechanisms [34]. More broadly, explainability contributes to trustworthy AI only when combined with socio-technical governance, auditability, responsibility, and institutional safeguards [35]. No isolated metric or evaluation layer is sufficient.

Table 4 presents the umbrella synthesis of explainability evaluation evidence, including cross-dimensional tensions, evaluation implications, remaining gaps, and future evaluation needs.

Table 4. Umbrella Synthesis of Explainability Evaluation Evidence: Fidelity Versus Plausibility, Faithfulness Versus Comprehensibility, Stability Versus Context Sensitivity, Understanding Versus Cognitive Load, Trust Versus Trustworthiness, Trust Calibration Versus Persuasion, Decision Usefulness Versus Automation Bias, Technical Metrics Versus Human Outcomes, and Future Evaluation Priorities

Synthesis theme	Evaluation dimensions involved	Review-level evidence pattern	Tension or trade-off	Evaluation implication	Remaining evidence gap	Future evaluation need
Fidelity versus plausibility	Faithfulness and human judgment	Usually evaluated separately	Convincing accounts may not reflect model behavior	Measure both independently	Few joint tests	Dual technical–human protocols
Faithfulness versus comprehensibility	Technical validity and accessibility	Faithful detail may be complex	Accuracy versus processing ease	Document simplification and information loss	Weak translation evidence	Layered explanations with fidelity checks
Stability versus context sensitivity	Stability, robustness, and sensitivity	Perturbation definitions vary	Invariance versus necessary response	Define semantically relevant changes	No standard contextual protocol	Joint stability–sensitivity assessment
Simplicity versus completeness	Sparsity, coverage, and cognitive load	Simple outputs are frequently preferred	Ease versus omitted information	Evaluate task-relevant sufficiency	Unknown optimal complexity	Adaptive detail and progressive disclosure
Understanding versus cognitive load	Comprehension and mental models	Self-report dominates	Detail may aid coverage but impair processing	Use objective and workload measures	Limited retention evidence	Longitudinal mental-model testing
Trust versus trustworthiness	Human attitude and system assurance	Constructs are frequently blurred	Confidence may be unwarranted	Evaluate trust alongside system evidence	Few integrated studies	Joint trustworthiness–trust evaluation
Trust calibration versus persuasion	Calibration and influence	Explanations can increase acceptance	Warranted confidence versus persuasion	Condition effects on AI correctness	Limited persuasion audits	Wrong-advice and misleading-explanation tests

Decision usefulness versus automation bias	Task support and overreliance	Benefits depend on task and error context	Assistance versus cognitive substitution	Measure verification and error correction	Artificial tasks dominate	Realistic decision studies
Technical metrics versus human outcomes	Fidelity, comprehension, trust, and decisions	Evaluation layers remain weakly connected	Model validity versus practical consequence	Use integrated, staged assessment	Sparse causal linkage	Multilevel evaluation designs
Short-term versus longitudinal evidence	Trust, learning, and reliance	Most user studies are brief	Immediate response versus adaptation	Interpret immediate effects cautiously	Limited repeated-use evidence	Longitudinal and field evaluation
No single-metric sufficiency boundary	All dimensions	Evidence supports heterogeneous measures	Comparability versus oversimplification	Use purpose-specific portfolios	No universal validated score	Construct-to-measure mapping

Figure 2 presents the umbrella synthesis model of explainability evaluation, linking technical explanation properties, human comprehension, decision usefulness, trust calibration, reliance behavior, evaluation validity, and future evaluation priorities.

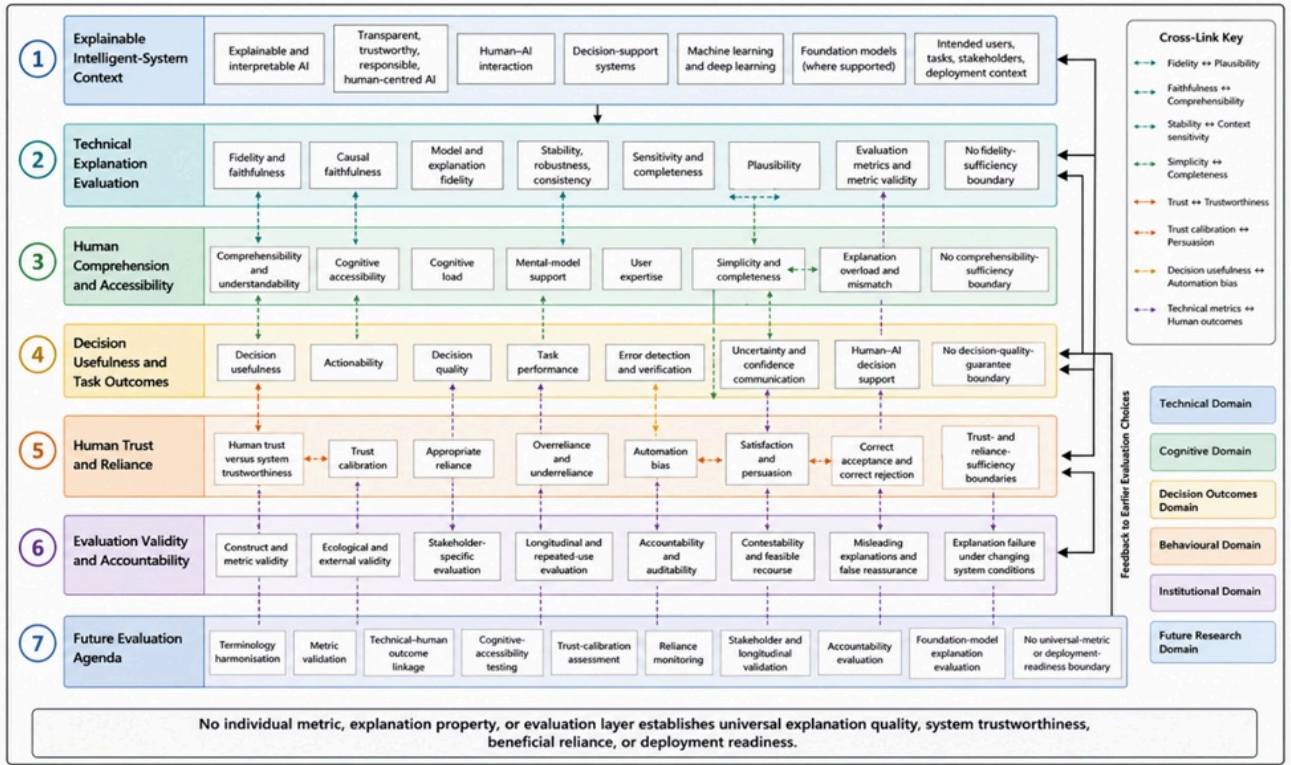


Figure 2. Umbrella Synthesis Model of Explainability Evaluation Outcomes

Alt-text description:

A conceptual umbrella synthesis model showing explainability evaluation across technical properties, human comprehension, cognitive accessibility, decision usefulness, trust calibration, reliance, evaluation validity, accountability, and future evaluation gaps.

Evidence Gaps and Future Evaluation Agenda

The first priority is construct and metric validity. Fragmented terminology permits fidelity, faithfulness, plausibility, comprehensibility, satisfaction, trust, and reliance to be treated as interchangeable. Future research should define each construct before selecting a measure, state the inference that the measure can support, and test convergent, discriminant, and predictive validity. Proposed metric portfolios should remain purpose-specific rather than becoming universal explanation scores.

A second priority is stronger linkage between technical properties and human outcomes. Evaluations should examine whether variations in fidelity, stability, completeness, uncertainty, or explanation complexity alter objective understanding, cognitive load, error detection, decision quality, and reliance. Stability protocols should distinguish irrelevant perturbations from contextually meaningful change. Comprehensibility studies should include lay users, technical experts, domain professionals, and affected stakeholders rather than assuming that one representation serves all audiences.

A third priority concerns realistic and longitudinal evaluation. Short-term experiments should be complemented by repeated-use studies that assess learning, mental-model development, changing trust, reliance adaptation, system updates, distribution shift, and accumulated error. Decision studies should include correct and incorrect AI advice, meaningful uncertainty, consequences for disagreement, and opportunities for verification. Accessibility, workflow fit, explanation overload, underreliance, overreliance, and automation bias should be treated as primary evaluation concerns rather than incidental observations.

The final priority is institutional validity. Explanations used in consequential systems should be evaluated for whether they support auditability, contestability, feasible recourse, correction, and accountable decision processes. Foundation-model evaluation should distinguish generated rationales from evidence about model mechanisms and should test fluent explanations for misleading confidence. These directions are proposed evaluation priorities grounded in recurring review-level gaps; they do not constitute a validated universal framework or establish legal compliance, operational safety, certification, or deployment readiness.

Conclusion

This umbrella review synthesizes what explainable-AI evaluations measure and what they do not establish across fidelity, faithfulness, stability, robustness, comprehensibility, cognitive accessibility, decision usefulness, human trust, trust calibration, and reliance. The evidence shows that explanation evaluation is multidimensional: technical correspondence does not guarantee understanding, perceived clarity does not prove faithfulness, trust does not equal trustworthiness, and greater reliance does not necessarily improve decisions. Stronger evaluation should connect technical properties to objective comprehension, realistic decision contexts, uncertainty, verification behavior, calibrated reliance, and institutional opportunities for accountability and contestability. The contribution is therefore not a universal explanation metric, but a review-level synthesis that clarifies evidentiary boundaries and supports more valid, stakeholder-specific, longitudinal, and socio-technical evaluation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Sep 2025

Revised: 03 Nov 2025

Accepted: 28 Nov 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.
- Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.
<https://doi.org/10.1145/3236009>.
- Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-60.
<https://doi.org/10.1109/ACCESS.2018.2870052>.
- Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.
- Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst*. 2021;11(3-4):1-45.
<https://doi.org/10.1145/3387166>.
- Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89-106.
<https://doi.org/10.1016/j.inffus.2021.05.009>.
- Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv*. 2023;55(13s):1-42.

<https://doi.org/10.1145/3583558>.

Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion*. 2023;99:101805.

<https://doi.org/10.1016/j.inffus.2023.101805>.

Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl Based Syst*. 2023;263:110273.

<https://doi.org/10.1016/j.knosys.2023.110273>.

Rong Y, Leemann T, Nguyen TT, Fiedler L, Qian P, Unhelkar V, et al. Towards human-centered explainable AI: a survey of user studies for model explanations. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(4):2104-22.

<https://doi.org/10.1109/TPAMI.2023.3331846>.

Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. *Artif Intell Rev*. 2022;55(5):3503-68.

<https://doi.org/10.1007/s10462-021-10088-y>.

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf Fusion*. 2024;106:102301.

<https://doi.org/10.1016/j.inffus.2024.102301>.

Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: a review of methods and applications. *Proc IEEE*. 2021;109(3):247-78.

<https://doi.org/10.1109/JPROC.2021.3060483>.

Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. *Data Min Knowl Discov*. 2023;37(5):1719-78.

<https://doi.org/10.1007/s10618-023-00933-9>.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.

<https://doi.org/10.1038/s42256-019-0048-x>.

Zhang Y, Tiño P, Leonardis A, Tang K. A survey on neural network interpretability. *IEEE Trans Emerg Top Comput Intell*. 2021;5(5):726-42.

<https://doi.org/10.1109/TETCI.2021.3100641>.

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750.

[https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).

Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. 2021;32(11):4793-813.

<https://doi.org/10.1109/TNNLS.2020.3027314>.

Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.

<https://doi.org/10.1016/j.jbi.2020.103655>.

Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.

<https://doi.org/10.1186/s12911-020-01332-6>.

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell.* 2021;296:103473.
<https://doi.org/10.1016/j.artint.2021.103473>.

Schoonderwoerd TAJ, Jorritsma W, Neerincx MA, van den Bosch K. Human-centered XAI: developing design patterns for explanations of clinical decision support systems. *Int J Hum Comput Stud.* 2021;154:1-25.
<https://doi.org/10.1016/j.ijhcs.2021.102684>.

Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann.* 2020;14(2):627-60.
<https://doi.org/10.5465/annals.2018.0057>.

Bach TA, Khan A, Hallock H, Beltrão G, Sousa S. A systematic literature review of user trust in AI-enabled systems: an HCI perspective. *Int J Hum Comput Interact.* 2024;40(5):1251-66.
<https://doi.org/10.1080/10447318.2022.2138826>.

Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI.* 2024;3:e53207.
<https://doi.org/10.2196/53207>.

Reinhard P, Li MM, Peters C, Leimeister JM. Effects of explanations in human-AI interaction: a systematic review and framework for future research on explainable AI. *Inf Syst Front.* 2026;28:1233-66.
<https://doi.org/10.1007/s10796-026-10716-4>.

Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. *Int J Hum Comput Stud.* 2023;169:102941.
<https://doi.org/10.1016/j.ijhcs.2022.102941>.

Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Comput Human Behav.* 2023;139:107539.
<https://doi.org/10.1016/j.chb.2022.107539>.

Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. *Proc ACM Hum Comput Interact.* 2023;7(CSCW1):1-38.
<https://doi.org/10.1145/3579605>.

Romeo G, Conti D. Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI Soc.* 2026;41(1):259-78.
<https://doi.org/10.1007/s00146-025-02422-7>.

Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: from principles to practices. *ACM Comput Surv.* 2023;55(9):1-46.
<https://doi.org/10.1145/3555803>.

Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Comput Surv.* 2022;55(5):1-29.
<https://doi.org/10.1145/3527848>.

Stepin I, Alonso JM, Catalá A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access.* 2021;9:11974-12001.
<https://doi.org/10.1109/ACCESS.2021.3051315>.

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. *ACM Trans Intell Syst Technol.* 2024;15(2):1-38.
<https://doi.org/10.1145/3639372>.

Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.
<https://doi.org/10.1007/s12525-020-00441-4>.