

REVIEW

Open access

Beyond Opening the Black Box—A Critical Review of Explainability for Human Understanding, Decision Usefulness, Trust Calibration, and Accountability

Thomas Andersen^{1*}, Sofia Lindqvist², David Okoro³

Abstract

Explainability has become central to discussions of trustworthy and responsible intelligent systems, yet it is frequently reduced to the technical production of information about model structure, features, rules, or predictions. This critical review examines why opening a computational black box is not equivalent to producing an explanation that people can understand or use responsibly. The review synthesises conceptual, technical, human-centred, decisional, and accountability-oriented scholarship to evaluate the shift from model interpretability toward meaningful explanation. It distinguishes technical properties such as transparency, fidelity, plausibility, completeness, stability, and local or global scope from human outcomes such as comprehension, cognitive accessibility, mental-model development, decision usefulness, calibrated trust, and appropriate reliance. It further examines how explanations may produce overreliance, underreliance, automation bias, false reassurance, cognitive overload, or explanation mismatch when their content and presentation are not aligned with users, tasks, uncertainty, and decision consequences. From an accountability perspective, explanations may contribute to auditability, traceability, contestability, and recourse, but they cannot independently assign responsibility, provide institutional authority, or guarantee effective remedies. The main contribution is a human-centred critical synthesis that treats explainability as a socio-technical achievement rather than a model output. Meaningful explanation requires technical credibility, cognitive accessibility, contextual relevance, decision-sensitive evaluation, uncertainty communication, opportunities for challenge, and governance mechanisms capable of acting on explanatory evidence. The review therefore rejects universal explanation-sufficiency claims and identifies conditions under which explanations may assist understanding and responsible decision use without being treated as proof of trustworthiness, accountability, fairness, safety, or deployment readiness.

Keywords Explainable AI, Interpretability, Accountability, Trust calibration, Human understanding, Decision usefulness

*Correspondence:

Thomas Andersen

thomas.andersen@di.ku.dk

¹ Department of Explainable AI and Human-Computer Interaction, Faculty of Engineering, University of Copenhagen, Copenhagen, Denmark

² Department of AI and Society, Stockholm University, Stockholm, Sweden

³ Department of Computer Science, University of Lagos, Lagos, Nigeria

Introduction

Explainability is commonly presented as a response to the opacity of machine-learning and intelligent decision systems. However, an explanation is not merely a technical description extracted from a model. Human explanations are selective, contrastive, socially situated, and shaped by the questions, expectations, and knowledge of their recipients [1]. This

insight challenges the assumption that disclosing internal variables, feature rankings, rules, or visualisations is sufficient to explain a decision. In human-centred intelligent systems, explanation quality must therefore be assessed not only by what information is produced, but also by whether that information supports accurate understanding for a specified person and purpose.

The explainable-AI literature encompasses intrinsic interpretability, post-hoc methods, local and global explanations, feature attribution, counterfactuals, examples, rules, and model-agnostic surrogates. It also connects explanation to responsible-AI concerns involving fairness, accountability, transparency, trust, and human oversight [2]. These dimensions are related but not interchangeable. A system may expose technically informative model behaviour while remaining inaccessible to a lay user, irrelevant to a professional decision, misleading about causation, or disconnected from institutional responsibility. Explainability should therefore be treated as a multidimensional property whose adequacy varies across stakeholders, risks, and intended uses.

The metaphor of opening the black box has helped motivate research into methods that reveal or approximate model behaviour. Nevertheless, the metaphor can imply that opacity is a single barrier and that exposing internal content resolves the problem. Existing surveys instead show a heterogeneous field with competing definitions, purposes, methods, and evaluation practices [3]. Some explanations describe global model patterns, whereas others address a single prediction. Some expose learned associations, while others provide examples, simplified rules, or hypothetical changes. These outputs answer different questions and should not be interpreted as interchangeable evidence of understanding.

Explanation problems differ according to the model being examined, the type of output required, the scope of the explanation, and the practical question faced by the user [4]. A technically faithful local attribution may be valuable for debugging but inadequate for contesting an institutional decision. A concise rule may be readable but omit important interactions. A counterfactual may identify a mathematical route to another prediction while offering no feasible route to a different real-world outcome. This article critically reviews explainability beyond technical disclosure by linking model interpretability to human understanding, cognitive accessibility, decision usefulness, trust calibration, appropriate reliance, accountability, contestability, and explanation failure.

From Model Interpretability to Meaningful Explanation

The distinction between intrinsic interpretability and post-hoc explainability is especially important in consequential applications. Intrinsically interpretable models expose decision logic through representations intended to remain directly inspectable, whereas post-hoc methods approximate or describe the behaviour of an already trained model. In high-stakes settings, post-hoc explanations should not automatically substitute for interpretable modelling when an adequately performing interpretable alternative is available [5]. The concern is not that every black-box explanation is useless, but that an approximate explanatory layer may conceal errors, unstable relationships, or decision rules that the explanation does not faithfully reproduce.

Meaningful explanation is consequently a pragmatic rather than purely structural achievement. Its value depends on what a stakeholder needs to understand, why that understanding matters, and what the stakeholder can do with the resulting information [6]. Developers may seek debugging evidence, domain experts may need uncertainty and causal relevance, operators may require error indicators, and affected individuals may need reasons that support correction or challenge. The same technical output can therefore be informative in one context and inadequate in another. Explanation purpose must precede method selection rather than being inferred after an explanation has been generated.

Interpretability itself has multiple meanings, including simplicity, decomposability, predictability, inspectability, intelligibility, and the possibility of constructing a usable account of system behaviour [7]. These meanings should not be collapsed. A model may be transparent at the level of mathematical structure but incomprehensible to its intended user. Conversely, a non-technical explanation may be understandable while providing only a partial or potentially inaccurate representation of

the model. Comprehensibility concerns whether information can be processed, while human understanding additionally concerns whether a person forms an accurate and practically relevant mental model.

Interpretability is best treated as an instrumental property whose importance depends on the epistemic, practical, or normative objective being pursued [8]. It does not independently establish accuracy, fairness, accountability, safety, or trustworthiness. Meaningful explanation requires a defined audience, a decision context, an explanation purpose, an acceptable level of detail, and evidence that the presentation supports the intended form of understanding. The critical boundary is therefore clear: model interpretability may contribute to meaningful explanation, but it cannot be treated as equivalent to human understanding or as proof that an intelligent system is responsibly designed or used.

Technical Explainability and Its Conceptual Limitations

Technical explainability includes intrinsic models and post-hoc approaches such as feature attribution, saliency visualisation, counterfactuals, examples, extracted rules, surrogate models, and local or global summaries. In safety-sensitive fields, these approaches differ substantially in what they reveal, how they are generated, and what users may reasonably infer from them [9]. Feature attribution describes sensitivity or contribution under particular assumptions, while saliency highlights influential input regions. Neither output necessarily identifies a causal mechanism, a complete decision process, or an intervention that will remain valid under changed conditions.

Methods for explaining deep neural networks include gradient-based analysis, relevance propagation, perturbation techniques, representation analysis, and surrogate approximation. These techniques can reveal selected aspects of learned behaviour, identify correlations, and assist model inspection, but they do not reproduce an exhaustive account of internal reasoning [10]. A visually coherent heat map may be unstable, and a concise surrogate may fit only a narrow neighbourhood. Natural-language explanations may improve accessibility, yet fluent wording can create an impression of certainty or causal insight that exceeds the evidence available from the model.

Explanation quality is multidimensional. Fidelity concerns whether an explanation corresponds to the behaviour of the target model; plausibility concerns whether it appears reasonable to a person; stability concerns whether materially similar inputs receive consistent explanations; and completeness concerns how much relevant model behaviour is represented [11]. These properties can conflict. A highly faithful account may be too complex for practical use, while a simple and plausible explanation may omit interactions or exceptions. Technical evaluation must therefore avoid treating any single property as a universal indicator of explanation quality.

Evaluation also requires attention to compactness, consistency, sensitivity, representativeness, robustness, and the relationship between explanation scope and the claim being made. Reviews of XAI evaluation show that many explanations are assessed through limited demonstrations or isolated metrics rather than comprehensive tests of their technical and human-relevant properties [12]. Local explanations should not be generalised to global behaviour without evidence, and global summaries should state what is compressed or omitted. Counterfactuals require feasibility and causal scrutiny, while example-based explanations require valid similarity measures and privacy-sensitive presentation.

The principal conceptual limitation is that technical explanations remain representations produced for particular purposes under particular assumptions. They can assist inspection, debugging, communication, or decision support, but their existence does not prove that a user understands the system, relies on it appropriately, or can hold an institution accountable. Technical evaluation should establish scope, fidelity, stability, completeness, uncertainty, and relevant failure conditions before human usefulness is inferred. **Table 1** summarises technical explainability approaches and their conceptual limitations when interpreted as meaningful human explanation.

Table 1. Technical Explainability Approaches and Conceptual Limitations: Model Transparency, Intrinsic Interpretability, Post-Hoc Explanation, Feature Attribution, Saliency, Counterfactuals, Example-Based Explanations, Rule-Based

Explanations, Surrogate Models, Local and Global Explanations, Fidelity, Plausibility, Stability, Completeness, and Technical-Sufficiency Boundaries

Explainability approach or concept	Core technical question	Explanation function	Conceptual limitation	Potential misinterpretation	Evaluation need	Critical boundary
Model transparency	Can relevant model structures or operations be inspected?	Exposes internal components or logic	Visibility may remain cognitively inaccessible	Visible operation is assumed to be understood	Inspectability and user-comprehension assessment	Transparency does not guarantee understanding
Intrinsic interpretability	Can prediction logic be represented directly?	Reduces reliance on post-hoc approximation	Simplicity is domain- and user-dependent	Interpretable form is treated as valid or trustworthy	Domain validity, error and usability testing	Interpretability does not establish correctness
Post-hoc explanation	Can black-box behaviour be described after prediction?	Produces an explanatory approximation	May diverge from actual model behaviour	Approximation is presented as internal reasoning	Faithfulness, robustness and scope testing	Transparency does not explain, not the other way around
Feature attribution	Which features influenced an output?	Ranks or scores input contribution	Depends on baselines, perturbations and correlation	Association is interpreted as causation	Sensitivity, baseline and correlation analysis	Attribution is not a complete picture
Saliency explanation	Which input regions appear influential?	Provides visual localisation	May be unstable or visually persuasive	Highlighted regions are treated as causal evidence	Sanity checks and perturbation testing	Visual plausibility is not sufficient
Counterfactual explanation	What change would alter the prediction?	Provides contrastive information	Minimal change may be infeasible or non-causal	Prediction change is treated as guaranteed recourse	Feasibility, causality and fairness testing	Counterfactuals differ from reality, not effective recourse
Example-based explanation	Which cases resemble the current case?	Supports reasoning through comparison	Similarity may be technically but not semantically valid	Similar examples are treated as equivalent cases	Similarity, representativeness and privacy review	Resemblance does not establish justification
Rule-based explanation	Can behaviour be represented as conditions and rules?	Produces symbolic decision logic	Extracted rules may be approximate or complex	Rule form is assumed to be understandable	Fidelity, coverage and complexity testing	Symbolic representation does not guarantee comprehensive understanding
Surrogate model	Can a simpler model approximate the black box?	Provides local or global approximation	Approximation error may be hidden	Surrogate logic is treated as original logic	Local and global fidelity assessment	Surrogate requires specific scope

Local explanation	Why was this output produced?	Explains behaviour near one instance	May not generalise beyond the selected neighbourhood	Local behaviour is treated as global behaviour	Neighbourhood and stability testing	Local v not g val
Global explanation	How does the model behave overall?	Summarises general relationships	Compression omits exceptions and interactions	Summary is treated as complete logic	Coverage and subgroup analysis	Glo explan ren sele
Causal explanation	What mechanism produced or could change an outcome?	Supports intervention-oriented understanding	Predictive data may not identify causation	Correlation is labelled causal	Causal assumptions and intervention validation	Prec explan not c expla
Natural-language explanation	Can reasons be expressed conversationally?	Improves linguistic accessibility	Fluency may conceal incompleteness or inaccuracy	Persuasive language is treated as faithful reasoning	Factuality, fidelity and uncertainty testing	Fluenc model
Explanation fidelity	Does the explanation represent model behaviour accurately?	Measures correspondence with the target model	High fidelity may remain unusable	Fidelity is treated as sufficient quality	Method-specific validity testing	Tech correc not h usefu
Explanation plausibility	Does the explanation appear reasonable to users?	Supports cognitive fit	May diverge from actual model behaviour	Convincing content is treated as true	Joint plausibility–fidelity assessment	Plaus cannot fid
Explanation stability	Does the explanation persist under non-material change?	Identifies sensitivity	Stable explanations may still be wrong	A single display is treated as dependable	Repeated and perturbation-based testing	O expla can esta sta
Explanation completeness	How much relevant behaviour is represented?	Assesses omissions and coverage	Completeness can conflict with accessibility	Concision is treated as comprehensiveness	Coverage and omission analysis	Sim req declar
Technical-sufficiency boundary	Can technical output establish meaningful explanation by itself?	Defines the critical limit	Human, contextual and governance conditions remain untested	Method availability is equated with responsible use	Cross-dimensional evaluation	No te me guar unders or rea

Human Understanding and Cognitive Accessibility

Human-centred explanation begins by identifying who must understand what, for which task, and under what constraints. Explanation design and evaluation should be aligned with distinct user groups and system objectives rather than assuming a generic recipient [13]. A developer investigating model errors, a clinician reviewing a recommendation, an auditor examining traceability, and an individual challenging an adverse outcome require different evidence. Lay users may need plain language and concrete consequences, whereas expert users may require technical detail, uncertainty, provenance, and access to alternative explanations.

Stakeholders differ not only in expertise but also in authority, vulnerability, time, incentives, and access to supporting evidence. Their explanatory needs may conflict: concise information for an affected person may be insufficient for an auditor, while extensive technical documentation may be unusable during a time-sensitive decision [14]. Cognitive accessibility therefore includes vocabulary, modality, sequencing, information density, disability access, and the opportunity to ask follow-up questions. Explanation personalisation may improve fit, but it should not conceal relevant information or become a means of persuasion.

Domain expertise is particularly important because explanation usefulness depends on whether technical information maps to legitimate concepts and workflows. Human-centred clinical-design research illustrates the value of eliciting explanation requirements from intended professional users before designing explanation interfaces [15]. Similar reasoning applies in other consequential domains. Feature importance alone may not answer whether evidence is clinically plausible, financially actionable, operationally safe, or institutionally contestable. Domain-specific explanation must clarify assumptions, uncertainty, scope, and the relationship between model output and professional judgment.

Technical explanation outputs become meaningful only after passing through user expertise, cognitive accessibility, mental-model formation, task context, uncertainty interpretation, and decision need. Human-centred studies remain heterogeneous in their measures of understanding, usability, trust, and task performance, which limits universal conclusions about explanation effectiveness [16]. Direct evaluation should therefore examine whether users can identify system capabilities, limitations, errors, relevant alternatives, and appropriate actions rather than relying only on satisfaction or perceived clarity. Figure 1 illustrates the pathway from technical explanation output to meaningful human explanation, showing how user expertise, cognitive accessibility, context, uncertainty, and decision needs shape understanding and usefulness.

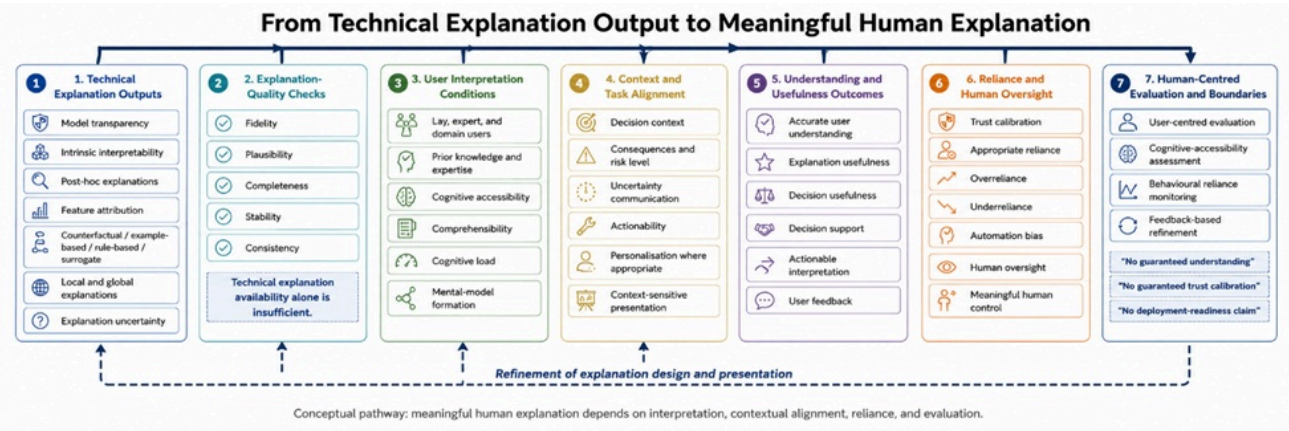


Figure 1. From Technical Explanation Output to Meaningful Human Explanation

Alt-text description:

A conceptual pathway diagram showing technical explanation outputs moving through user expertise, cognitive accessibility, mental models, task context, uncertainty interpretation, decision usefulness, feedback, and human-centred

evaluation.

Cognitive accessibility and comprehensibility are necessary but remain insufficient indicators of understanding. A user may understand the literal content of an explanation while forming an inaccurate mental model of system capability, uncertainty, or causation. Conversely, an explanation can be technically complex yet useful to a qualified expert when it supports error detection or comparison with domain evidence. Human-centred evaluation should therefore combine comprehension, cognitive-load, mental-model, task, accessibility, and behavioural measures. The governing boundary is that providing an explanation does not guarantee understanding; meaningful explanation must be demonstrated for intended users under defined decision conditions.

Decision Usefulness, Trust Calibration, and Reliance

Trust in artificial intelligence is not a single attitude produced by explanation availability. It includes cognitive judgments about competence, predictability, and reliability as well as affective responses shaped by task, interface, prior experience, and perceived risk [17]. An explanation may increase confidence because it appears coherent, but increased confidence is not necessarily calibrated confidence. Trustworthiness concerns properties of the system and its operation, whereas trust concerns a user's perception. Explainable systems should therefore be evaluated by whether perceptions and behaviour correspond appropriately to demonstrated capability.

The more defensible objective is appropriate reliance rather than trust maximisation [18]. Appropriate reliance occurs when users accept useful recommendations under conditions in which the system is sufficiently capable and reject, verify, or escalate outputs when limitations or warning signs are present. Overreliance involves unwarranted acceptance, while underreliance involves rejecting beneficial assistance without adequate reason. Explanations may support calibration by communicating limitations, uncertainty, alternatives, and relevant evidence, but they can also legitimise an output merely by making it appear reasoned.

Perceived explainability and causability may influence trust, acceptance, and users' sense that a system is understandable [19]. Such effects should be interpreted cautiously because perceived clarity does not independently establish explanation fidelity, causal validity, or improved decision quality. A user can understand the literal content of an explanation yet misunderstand the model's competence, the source of uncertainty, or the consequences of acting on its recommendation. Confidence communication should therefore distinguish prediction confidence, explanation uncertainty, data limitations, and the possibility that the system is operating outside conditions represented in its validation.

Evidence from high-risk decision tasks indicates that explanations do not produce uniform improvements in trust or behaviour [20]. Their effects depend on explanation design, user preparation, error visibility, time pressure, and the decision consequences involved. In some settings, explanations may encourage verification; in others, they may intensify automation bias by supplying a plausible justification for an incorrect recommendation. Human oversight is meaningful only when reviewers possess sufficient expertise, time, authority, alternative evidence, and an operational capacity to reject or escalate the automated output.

Interactive explanation design can support feedback and clarification, yet interaction may also produce overload, selective attention, or calibration errors when users receive information that does not match their needs [21]. Reliance should therefore be monitored behaviourally and longitudinally rather than inferred from satisfaction or trust ratings alone. Evaluation should examine acceptance of correct and incorrect advice, independent error detection, override behaviour, uncertainty interpretation, and changes in reliance after system failures. Providing an explanation does not guarantee calibrated trust, and understanding an explanation does not guarantee a better decision.

Accountability, Contestability, and Explanation Failure

Accountability requires more than visibility into a model. Transparency can reveal selected components while leaving users unable to understand organisational relationships, data practices, temporal dependencies, or the distribution of authority surrounding a decision [22]. The black-box metaphor can consequently transfer opacity from the predictive model to the explanation method, interface, documentation system, or deploying institution. An explanation becomes relevant to accountability only when it forms part of a process capable of identifying responsibility, reviewing evidence, correcting errors, and responding to harm.

Technical disclosure alone cannot restore accountability without institutional mechanisms such as audit access, traceability, documented responsibilities, review authority, corrective procedures, and consequences for failure [23]. Auditability requires qualified reviewers to inspect evidence beyond a simplified user explanation. Traceability requires outputs to be connected to data sources, model versions, explanation methods, operators, and interventions. Documentation must remain current, accessible, and purpose-specific. Otherwise, explanation can become ceremonial disclosure that creates an appearance of oversight without enabling effective scrutiny.

Responsibility attribution is also relational. Explanations may clarify how an output was generated or used, but they cannot independently determine who had the duty, authority, knowledge, or capacity to prevent harm [24]. Developers, data providers, deployers, managers, professional users, and institutional decision-makers may each control different parts of the system. Accountability-oriented explanation must therefore be connected to explicit role allocation, escalation channels, and human control rather than allowing responsibility to be displaced onto either the model or the person receiving its recommendation.

Transparency requirements must be sensitive to audience, purpose, context, and competing concerns [25]. Information that supports technical auditing may be inaccessible to an affected person, while unrestricted disclosure can create privacy or security risks. Example-based explanations may reveal sensitive records; detailed system information may facilitate manipulation; and generic disclosure can obscure differential effects across groups. Fairness implications therefore concern not only model outcomes but also whether explanation quality, challenge burdens, and access to recourse are distributed equitably.

Contestability requires an affected person to question inputs, inferences, procedures, or outcomes through an accessible review process. Counterfactual information may clarify what model input would change a prediction, but it does not automatically provide feasible, causal, affordable, safe, or fair recourse [26]. Explanation failure includes misleading content, false reassurance, overload, mismatch, omitted uncertainty, inaccessible presentation, and non-actionable recommendations. **Table 2** maps human understanding, cognitive accessibility, decision usefulness, trust calibration, reliance, accountability, contestability, and explanation failure in human-centered explainable systems.

Table 2. Human-Centered Explainability Requirements and Failure Risks: Cognitive Accessibility, Mental Models, User Expertise, Decision Usefulness, Trust Calibration, Appropriate Reliance, Overreliance, Underreliance, Automation Bias, Accountability, Contestability, Recourse, Misleading Explanations, Explanation Overload, and Explanation Mismatch

Human-centered requirement or failure component	Core review question	User, decision, or governance context	Explanation role	Potential benefit or accountability output	Failure or misuse risk	Evaluation or oversight need
Human understanding	Does the user form an accurate and usable account?	Expertise, task and stakes vary	Connects technical information to a user question	Informed scrutiny	Persuasive but inaccurate mental model	Comprehension and transfer testing

Cognitive accessibility	Can the user perceive and process the explanation?	Language, disability, workload and literacy matter	Adapts format, vocabulary and sequencing	More equitable access	Exclusion through jargon or inaccessible presentation	Accessibility and usability testing
Mental-model support	Does the explanation clarify capabilities and limits?	Prior beliefs influence interpretation	Supports capability learning	Improved error anticipation	Entrenched misconceptions	Pre/post mental-model assessment
User expertise	What knowledge does the explanation assume?	Lay, technical and domain experts differ	Adjusts abstraction and evidence depth	Better stakeholder fit	Oversimplification or overload	Stratified evaluation
Decision usefulness	Does the explanation assist the defined task?	Decision purpose and consequence differ	Provides relevant evidence and alternatives	Better scrutiny or option comparison	Distraction or irrelevant information	Task and consequence evaluation
Trust calibration	Does confidence track demonstrated capability?	Reliance changes across tasks and failures	Communicates limitations and uncertainty	Better aligned use	Persuasion without calibration	Behavioural calibration assessment
Appropriate reliance	Do users accept and reject outputs appropriately?	Requires alternatives and authority	Supports verification and override	Safer human-AI collaboration	Deference or unjustified rejection	Correct- and incorrect-advice testing
Overreliance	Are incorrect outputs accepted despite warning signs?	Time pressure and authority cues matter	May reveal limitations	Potential reduction in blind acceptance	Automation bias and false reassurance	Error-injection and override monitoring
Underreliance	Are useful outputs rejected without adequate reason?	Prior distrust or poor explanation may contribute	Clarifies validated strengths	Appropriate beneficial use	Disuse and missed benefit	Reliance and missed-opportunity measures
Automation bias	Does explanation suppress independent checking?	Recommendation and decision-support systems	Can prompt or displace verification	More structured review	Rubber-stamping	Independent-reasoning tasks

Accountability	Can reasons be connected to responsible actors and action?	Organisational authority is required	Supplies reviewable evidence	Correction and responsibility attribution	Explanation substitutes for responsibility	Role, authority and remedy review
Contestability	Can an affected person challenge a decision?	Rights, time, access and review channels matter	Provides understandable grounds	Review, correction or reversal	Inaccessible or non-disputable reasons	Procedural and outcome evaluation
Recourse	Is there a feasible route to another outcome?	Resources and causal constraints matter	Connects reasons to possible action	Consequential remedy	Infeasible or burdensome recommendation	Feasibility, fairness and causality testing
Misleading explanation	Does content imply unsupported causation or certainty?	Any user may be affected	Should qualify scope and evidence	Corrected interpretation	Plausible but inaccurate narrative	Fidelity and communication checks
False reassurance	Does explanation make a deficient system appear safe?	Especially serious in high-risk use	Communicates limits and warning signs	More cautious use	Unwarranted confidence	Independent validation and reliance testing
Explanation overload	Does information exceed cognitive or temporal capacity?	Time-sensitive and complex decisions	Prioritises and layers information	More manageable access	Skimming, avoidance or deference	Workload and recall testing
Explanation mismatch	Does the output answer the user's actual question?	Debugging, deciding and contesting differ	Aligns content with purpose	Relevant information	Technically correct but practically useless output	Requirements elicitation and task testing
Opacity transfer	Has opacity moved to the explanation or institution?	Explanation pipelines and organisations	Exposes transformations and provenance	End-to-end scrutiny	New explanatory black box	Traceability and audit access
Uncertainty omission	Are predictive and explanatory uncertainties absent?	High-risk or shifted contexts	Qualifies confidence and scope	Better risk interpretation	False certainty	Calibration and comprehension testing

Fairness implication	Are explanation access and burdens equitable?	Effects differ across stakeholders and groups	Supports identification of differential treatment	Corrective review	Unequal explanation quality or recourse burden	Subgroup and burden analysis
Privacy implication	Does explanation disclose sensitive information?	Examples, features and provenance can reveal data	Limits disclosure to relevant information	Proportionate transparency	Data leakage or attribute exposure	Privacy threat assessment
Security implication	Can explanatory detail facilitate attacks or gaming?	Public and role-based access differ	Supports controlled scrutiny	Secure accountability	Manipulation, extraction or gaming	Adversarial and misuse testing
Governance	Can the institution act on explanation evidence?	Policies, roles and remedies are required	Embeds explanation in oversight	Correction, learning and sanctions	Ceremonial compliance	Governance-effectiveness review

Critical Synthesis of Explainability Research

A proportionate approach should avoid demanding perfect introspective access from algorithms when comparable human decisions are themselves partly opaque, intuitive, or institutionally distributed [27]. This does not remove the need for algorithmic scrutiny. Rather, it redirects evaluation from an unattainable ideal of complete transparency toward evidence appropriate to risk, purpose, and decision authority. The central distinction is between interpretability as access to selected system properties and meaningful explanation as information that supports a legitimate human or institutional objective.

Technical fidelity and human plausibility represent separate but interacting dimensions. The distinction between machine explainability and human causability shows why an output can accurately reflect model behaviour yet fail to support a user’s causal understanding [28]. Conversely, a plausible account may align with human expectations while misrepresenting the model. Similar tensions arise between simplicity and completeness: concise explanations can reduce cognitive load but omit interactions, uncertainty, or exceptions, while comprehensive accounts may become unusable. Layered presentation may help, but its effectiveness requires evaluation rather than assumption.

The risks are especially clear where plausible post-hoc explanations are used to reassure professionals or affected persons about systems whose external validity, causal relevance, or distributional performance remains uncertain [29]. Explanation availability can obscure rather than resolve deficiencies when it implies that a prediction is clinically, socially, or institutionally justified. Understanding may then become confidence in a persuasive representation rather than an accurate appreciation of system capability. Explanation failure should therefore include false reassurance and contextual invalidity, not only low technical fidelity.

Across the literature, explainability challenges remain fragmented among model development, evaluation, human interaction, organisational deployment, and governance [30]. Standardisation can improve reporting of fidelity, stability, uncertainty, scope, and evaluation conditions, but a universal presentation format would ignore differences among users

and purposes. Transparency can also conflict with privacy and security, while contestability may require disclosure that is understandable and sufficiently detailed to support challenge. These tensions require explicit trade-off analysis rather than a single composite explanation score.

Explainability is only one component of trustworthy artificial intelligence and cannot replace fairness, robustness, privacy, security, accountability, or effective governance [31]. The purpose for which an explanation is requested should determine its design properties, evidentiary standard, intended user, and evaluation method [32]. The synthesis therefore rejects a universal explanation-sufficiency boundary: explanation availability is not equivalent to explanation usefulness, and no single method or metric can establish responsible use. **Table 3** presents the critical synthesis of explainability research, including tensions, failure modes, implications, remaining gaps, and research priorities.

Table 3. Critical Synthesis of Explainability Research Beyond Black-Box Opening: Interpretability Versus Meaningful Explanation, Technical Fidelity Versus Human Plausibility, Simplicity Versus Completeness, Understanding Versus Cognitive Load, Trust Calibration Versus Persuasion, Decision Usefulness Versus Automation Bias, Accountability Versus Opacity, Contestability Versus Inaccessibility, and Research Priorities

Critical synthesis theme	Dimensions involved	Core critical question	Evidence pattern	Tension or failure mode	Implication for explainable systems	Remaining gap
Interpretability versus meaningful explanation	Transparency, intelligibility, purpose	Does inspectable logic answer the user's question?	Technical access and stakeholder understanding are repeatedly distinguished	Interpretable output may be irrelevant or inaccessible	Specify audience, purpose and consequence first	Limited cross-stakeholder comparison
Technical fidelity versus human plausibility	Fidelity, plausibility, trust	Can an explanation be faithful and cognitively acceptable?	The properties are distinct and can diverge	Persuasive but unfaithful explanation	Test both dimensions and disclose trade-offs	Few validated joint measures
Simplicity versus completeness	Cognitive load, coverage, local/global scope	What may be omitted without distortion?	Compactness and completeness are separately valued	Simplicity conceals exceptions; completeness overwhelms	Use layered explanations with scope statements	No general compression threshold
Understanding versus cognitive load	Accessibility, mental models, expertise	Does more information improve understanding?	Needs vary by user and workflow	Overload, shallow processing or misunderstanding	Evaluate comprehension and workload together	Sparse inclusive and longitudinal evidence
Trust calibration versus persuasion	Trust, reliance, uncertainty	Does explanation align confidence with capability?	Explanation effects are mixed and context-dependent	Coherence raises trust without accuracy	Target appropriate reliance, not maximum trust	Limited behavioural calibration measures

Decision usefulness versus automation bias	Decision support, oversight, consequence	Does explanation assist judgment or legitimise automation?	Behavioural benefits are not automatic	Reduced independent verification	Test correct and incorrect advice use	Few operational studies
Accountability versus explanation opacity	Accountability, auditability, traceability	Can explanatory evidence support institutional action?	Transparency alone is insufficient	Opacity shifts to tools or organisations	Connect explanations to roles, audits and remedies	Limited workflow evaluation
Contestability versus inaccessible explanation	Contestability, recourse, accessibility	Can affected people use reasons to challenge outcomes?	Informational and consequential explanations differ	Technical reasons lack feasible challenge pathways	Co-design explanation and review mechanisms	Sparse studies with affected persons
Transparency versus privacy	Disclosure, examples, provenance	What information can be revealed proportionately?	Explainability and privacy can conflict	Sensitive information is exposed	Use role-based and purpose-limited access	Few integrated assessments
Standardisation versus context sensitivity	Metrics, stakeholder needs, governance	What should be standardised?	Recurring technical properties coexist with contextual usefulness	Universal format becomes irrelevant	Standardise reporting categories, not one explanation	Weak evidence on adaptation rules
Availability versus usefulness	Method output, comprehension, actionability	Does an available explanation fulfil its purpose?	Production is evaluated more often than practical value	Explanation is present but unusable	Treat availability as an input, not an outcome	Limited task and field validation
No universal sufficiency boundary	All dimensions	Can one method or metric prove responsible explanation?	Evidence shows user, purpose, risk and governance dependence	Composite scoring conceals missing evidence	Use purpose-specific evidence profiles	No agreed minimum evidence by context

Figure 2 presents the critical synthesis model of explainability research, linking technical explanation properties, human understanding, decision usefulness, trust calibration, accountability, contestability, explanation failure, and research priorities.

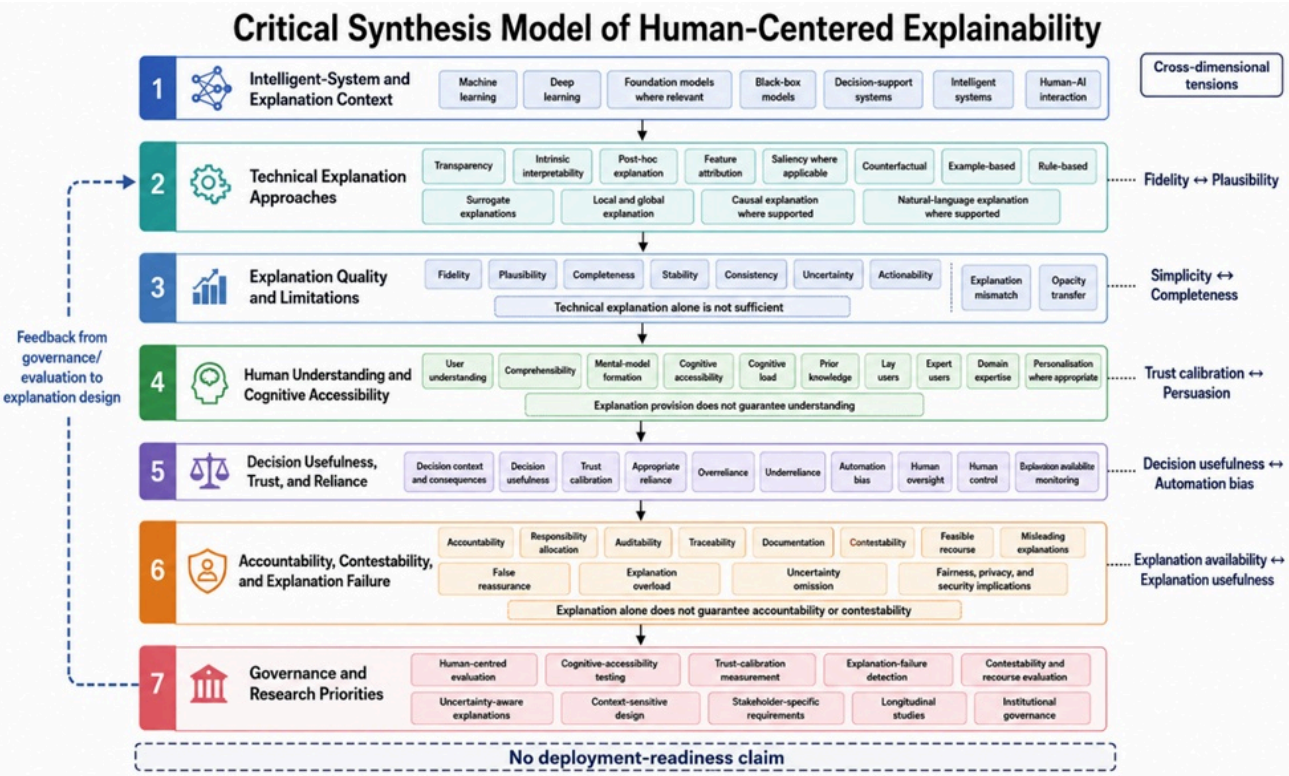


Figure 2. Critical Synthesis Model of Human-Centered Explainability

Alt-text description:

A conceptual synthesis model showing explainability as a system connecting technical explanation properties, user understanding, cognitive accessibility, decision usefulness, reliance, trust calibration, accountability, contestability, explanation failure, governance, and future research priorities.

Research Priorities for Human-Centered Explainable Systems

Future research should begin with human-centred explanation evaluation that defines the intended user, task, decision consequence, and expected form of understanding before selecting an explanation method. Cognitive-accessibility studies should examine language, disability access, cognitive load, expertise, mental-model formation, and retention rather than relying only on satisfaction or self-reported clarity [16]. Decision-usefulness validation should assess whether explanations improve error detection, evidence comparison, uncertainty interpretation, and appropriate action under both correct and incorrect system outputs.

Trust-calibration research should measure behaviour as well as attitudes. Studies should vary system competence, explanation accuracy, uncertainty, and failure visibility to determine whether users rely appropriately rather than simply reporting greater trust [18]. Reliance and automation-bias assessment should include override behaviour, independent verification, underuse, repeated interaction, and adaptation after observed errors [20]. Longitudinal studies are particularly important because initial novelty, explanation fatigue, habituation, and strategic system use may change outcomes over time.

Accountability-oriented explanation design should investigate how explanatory evidence moves through documentation, audit, review, correction, and remedy processes. Contestability and recourse evaluation should include affected persons

and examine whether reasons are accessible, whether challenges receive independent consideration, and whether proposed actions are feasible and causally relevant [26]. Explanation-failure detection should monitor instability, mismatch, overload, false reassurance, omitted uncertainty, privacy leakage, security exposure, and opacity transfer rather than treating technical fidelity as the only failure criterion.

Uncertainty-aware and context-sensitive explanations require methods that distinguish predictive uncertainty, data limitations, explanatory uncertainty, and out-of-distribution conditions. Stakeholder-specific designs should make necessary differences in depth and access auditable rather than allowing personalisation to hide relevant information. Foundation-model explanations require particular caution because fluent natural-language rationales may be plausible without being faithful accounts of internal computation. The broader research boundary is that no universal metric can establish meaningful explanation across users, purposes, risks, and institutions [32]. Evaluation should instead use purpose-specific evidence profiles while recognising that explainability remains only one component of trustworthy and responsible system governance [31].

Conclusion

This critical review moves beyond the metaphor of opening the black box by treating explainability as a human-centred and socio-technical achievement. Model transparency, intrinsic interpretability, post-hoc explanation, fidelity, plausibility, stability, and completeness provide important technical evidence, but they do not independently establish human understanding or responsible use. Meaningful explanation depends on cognitive accessibility, user expertise, mental-model support, task relevance, uncertainty communication, decision consequences, and empirical evaluation with intended users. It must also be assessed by whether it supports appropriate reliance rather than trust maximisation, preserves meaningful human oversight and control, and avoids overreliance, underreliance, automation bias, overload, mismatch, and false reassurance. In consequential systems, explanation becomes accountability-relevant only when it is connected to traceability, auditability, responsibility allocation, contestability, feasible recourse, and institutions capable of review and correction. The resulting synthesis rejects universal explanation-sufficiency claims and positions explanation as one conditional contribution to trustworthy intelligent systems. Future progress requires evaluation that links technical properties with human understanding, decision usefulness, calibrated reliance, explanation-failure detection, accountability, and contestability under clearly defined contexts.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Published online: 15 March 2025

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

- 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>.
- 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>.
- 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>.
- 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>.
- 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>.
- Páez A. The pragmatic turn in explainable artificial intelligence (XAI). *Minds Mach (Dordr).* 2019;29(3):441-59.
<https://doi.org/10.1007/s11023-019-09502-w>.
- Erasmus A, Brunet TDP, Fisher E. What is interpretability? *Philos Technol.* 2021;34(4):833-62.
<https://doi.org/10.1007/s13347-020-00435-2>.
- Krishnan M. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philos Technol.* 2020;33(3):487-502.
<https://doi.org/10.1007/s13347-019-00372-9>.
- 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>.
- 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>.
- 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>.

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>.

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;154:102684.
<https://doi.org/10.1016/j.ijhcs.2021.102684>.

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>.

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>.

Benda NC, Novak LL, Reale C, Ancker JS. Trust in AI: why we should be designing for appropriate reliance. *J Am Med Inform Assoc*. 2022;29(1):207-12.

Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int J Hum Comput Stud*. 2021;146:102551.
<https://doi.org/10.1016/j.ijhcs.2020.102551>.

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>.

Naiseh M, Al-Thani D, Jiang N, Ali R. Explainable recommendation: when design meets trust calibration. *World Wide Web*. 2021;24(5):1857-84.
<https://doi.org/10.1007/s11280-021-00916-0>.

Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc*. 2018;20(3):973-89.
<https://doi.org/10.1177/1461444816676645>.

de Laat PB. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philos Technol*. 2018;31(4):525-41.
<https://doi.org/10.1007/s13347-017-0293-z>.

Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci Eng Ethics*. 2020;26(4):2051-68.
<https://doi.org/10.1007/s11948-019-00146-8>.

Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc*. 2019;6(1):1-14.
<https://doi.org/10.1177/2053951719860542>.

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>.

Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: is there a double standard? *Philos Technol.* 2019;32(4):661-83.

<https://doi.org/10.1007/s13347-018-0330-6>.

Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2019;9(4):e1312.

<https://doi.org/10.1002/widm.1312>.

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).

Saeed W, Omlin CW. 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>.

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>.

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>.