

ORIGINAL RESEARCH

Open access

Which Explanation for Whom and When? A Stakeholder-Oriented Taxonomy of Global, Local, Counterfactual, Causal, Interactive, and Adaptive Explanations

Laura Mendez^{1*}, Hiroshi Sato², Anna Kowalczyk³

Abstract

Explainable artificial intelligence increasingly offers diverse ways to communicate how intelligent systems operate, yet explanation selection is still frequently driven by method availability or model compatibility rather than stakeholder need. This article develops a stakeholder-oriented taxonomy for determining which explanation type may be appropriate for whom, at which decision moment, and for what purpose. Explanation is treated as a multiform system capability encompassing different scopes, reasoning functions, presentation modes, interaction patterns, and quality requirements. The taxonomy distinguishes global explanations concerned with model-level behaviour from local explanations directed at individual predictions, while clarifying the roles and limitations of feature attribution, saliency, surrogate models, rule extraction, transparency, fidelity, stability, and completeness. It further differentiates counterfactual, contrastive, and causal explanations according to their contributions to alternative reasoning, actionability, contestability, recourse, and accountability. Example-based, case-based, prototype, rule-based, natural-language, interactive, dialogue-based, adaptive, and personalized explanations are positioned according to user expertise, cognitive accessibility, task demands, and interaction context. The proposed taxonomy connects explanation families with stakeholder roles, pre-decision, during-decision, and post-decision timing, explanation purpose, decision consequence, uncertainty, trust calibration, human oversight, and accountability requirements. Its central contribution is a structured selection logic that moves beyond technical method categories toward stakeholder–purpose–timing–risk alignment. The taxonomy can guide explanation design and evaluation but does not independently establish understanding, fairness, recourse, accountability, legal sufficiency, trustworthiness, or deployment readiness. Stakeholder-specific and decision-context-specific validation therefore remains necessary.

Keywords Explainable AI, Explanation taxonomy, Global explanation, Local explanation, Counterfactual explanation, Causal explanation

*Correspondence:

Laura Mendez
laura.mendez@ie.edu

¹ Department of Human-Centered AI and Explainability, IE Business School, Madrid, Spain

² Department of AI and Decision Science, Keio University, Tokyo, Japan

³ Department of Computer Science and AI, University of Warsaw, Warsaw, Poland

Introduction

Explanations in intelligent systems are not neutral descriptions transmitted from a model to a passive recipient. They are selective communications shaped by the question being asked, the information considered relevant, and the cognitive and social expectations of the person receiving them. Human explanations are often contrastive, audience-sensitive, and directed toward particular uncertainties rather than toward complete disclosure of every factor involved [1]. Consequently,

the central explainability problem is not merely whether a system can generate an explanation, but whether it can produce an appropriate explanation for a specific stakeholder and purpose.

Existing explainable artificial intelligence classifications provide valuable distinctions among intrinsic and post-hoc methods, model-specific and model-agnostic approaches, and local and global explanation techniques. These classifications establish an essential technical vocabulary, but technical form alone does not determine whether an explanation is cognitively accessible, decision-relevant, actionable, or suitable for accountability [2]. An explanation method may be technically compatible with a model while remaining mismatched to the stakeholder's question, expertise, decision authority, or exposure to harm.

Black-box explanation methods also differ in the object they explain, the information available to the explanation process, the scope of their claims, and the form of their outputs. Global summaries, local rules, feature relevance scores, representative examples, prototypes, counterfactuals, and surrogate models therefore answer different explanatory questions [3]. Treating these outputs as interchangeable can conceal important distinctions between describing general model behaviour, accounting for an individual prediction, identifying an alternative outcome, communicating a decision rule, and supporting scrutiny of a consequential decision.

This article develops an original stakeholder-oriented explainability taxonomy organized around four questions: which explanation is needed, by which stakeholder, at which decision moment, and under which purpose or risk condition? Stakeholder interests can include understanding, debugging, decision support, justification, contestation, recourse, oversight, or accountability, and these interests should not be reduced to fixed role-based assumptions [4]. The proposed taxonomy therefore connects explanation scope and reasoning function with stakeholder responsibility, expertise, timing, actionability, cognitive accessibility, uncertainty, trust calibration, and accountability demands.

Explanation as a Multiform System Capability

Explainability should be understood as a multiform system capability rather than a single property possessed uniformly by a model. Interpretability commonly concerns the extent to which model behaviour or reasoning can be understood, while transparency concerns access to relevant structures, processes, data relations, or decision logic. Intelligibility and comprehensibility refer more directly to whether an account can be cognitively processed by an intended audience. These concepts overlap, but they should not be treated as equivalent because they support different design objectives and evaluation approaches [5].

A multiform capability links explanation purpose, audience, presentation, and evaluation. Developers may require diagnostic information about model behaviour; domain experts may need evidence aligned with professional reasoning; operators may need timely indications of abnormal system behaviour; and affected individuals may need understandable reasons and meaningful routes for challenge. Explanation design should therefore connect the user group and intended goal with an appropriate explanation form and evaluation method rather than beginning with a preferred technical technique [6].

Interpretability is also multidimensional. A model or explanation may be descriptively accurate yet irrelevant to the stakeholder's task, practically useful yet incomplete, or simple enough to inspect while concealing important interactions. Predictive performance, descriptive fidelity, domain relevance, and human usability can diverge and may require separate evaluation [7]. Explanation quality should consequently be specified relative to the intended purpose instead of being represented as a single universal property.

Explanations may be static, such as a fixed feature-attribution display, or dynamic, such as an evolving dialogue that responds to follow-up questions. They may be uniform across users or adapted according to expertise, task state, cognitive needs, or interaction history. However, adaptive and personalized explanations introduce risks of stereotyping, selective omission, privacy intrusion, and inconsistent disclosure. A taxonomy can distinguish explanatory objects,

Global and Local Explanations

Model transparency can reduce dependence on post-hoc approximation when an intrinsically interpretable model is appropriate for the task. In consequential settings, a post-hoc explanation should not automatically be treated as an adequate substitute for a model whose decision logic can be directly examined [12]. Nevertheless, transparency to developers does not ensure accessibility to lay users, and a comprehensible local explanation does not establish global safety, fairness, or accountability. **Table 1** summarises global and local explanation types, their explanation scope, stakeholder needs, potential uses, failure risks, evaluation needs, and taxonomy roles.

Explanation type	Core explanation question	Scope of explanation	Likely stakeholder need	Potential use	Failure or misuse risk	Evaluation need	
Global explanation	How does the system generally behave?	Entire model or broad behavioural region	Developer, model builder, auditor, manager, domain expert	Inspection, monitoring, validation, risk review	Aggregate patterns may conceal local or subgroup failures	Fidelity, coverage, stability, subgroup analysis	Systemic error

Local explanation	Why was this output produced for this case?	Individual prediction or decision	Affected individual, decision maker, operator, domain expert	Case review, decision support, contestation	May be unstable, incomplete, or generalized beyond the case	Local fidelity, relevance, stability, uncertainty	e
Model-level explanation	What logic or relationships characterize the model?	Whole model	Developer, auditor, expert user	Debugging, documentation, governance	Simplification may omit interactions or exceptions	Completeness, complexity, global fidelity	Sy ar e
Prediction-level explanation	Which factors influenced this particular result?	Single output	End user, decision subject, operator	Individual decision understanding	Influence may be mistaken for causation or responsibility	Case fidelity, feature semantics, uncertainty	U e
Feature attribution	How much did each input contribute under the selected assumptions?	Local or aggregated global	Analyst, expert, affected person	Factor comparison, debugging, review	Dependence assumptions may distort reported contributions	Correctness, sensitivity, consistency	
Saliency explanation	Which input regions or elements were highlighted?	Usually local	Developer, expert, operator	Visual model inspection	Plausible visualization may be unfaithful or unstable	Perturbation tests, stability, expert relevance	Vis e
Surrogate explanation	What simpler model approximates the original system?	Global or local approximation	Developer, auditor, expert	Inspection and approximate rule extraction	Surrogate may diverge outside its validated scope	Scope-specific fidelity, coverage, stability	Ap e
Rule extraction	Which conditions appear to govern outputs?	Global or local	Expert, operator, auditor	Traceability and decision review	Rules may oversimplify probabilistic behaviour	Fidelity, exceptions, complexity, coverage	R e
Model transparency	Can internal logic be directly examined?	Intrinsic model	Developer, expert, auditor	Direct verification and documentation	Developer transparency may not equal user comprehensibility	Simulatability, decomposability, domain intelligibility	
Explanation scope	Is the account model-wide, regional, cohort-based,	Cross-cutting	All stakeholders	Preventing inappropriate generalization	Scope may be unstated or misunderstood	Explicit scope declaration and validation	

	or case-specific?						
Explanation granularity	How much detail is appropriate?	Cross-cutting	Depends on expertise and task	Layered communication	Too little omits relevant evidence; too much causes overload	Cognitive-load and task testing	P an
Fidelity, stability, and completeness	Does the explanation accurately, consistently, and sufficiently represent its stated object?	Global or local	All stakeholders	Technical and human evaluation	A plausible account may be inaccurate, unstable, or materially incomplete	Multidimensional, purpose-relative evaluation	C re
Global-explanation sufficiency boundary	Can a global account prove that an individual decision was appropriate?	Boundary condition	All stakeholders	Preventing overclaiming	System summaries may not resolve case-specific concerns	Instance-level and contextual validation	sy
Local-explanation sufficiency boundary	Can a local account establish system-wide safety, fairness, or accountability?	Boundary condition	All stakeholders	Preventing false reassurance	Persuasive case explanations may conceal systemic problems	Global auditing and outcome monitoring	ins

Counterfactual, Contrastive, and Causal Explanations

Counterfactual and contrastive explanations are closely related but serve distinguishable reasoning functions. A counterfactual typically asks what would need to change for a different output to occur, whereas a contrastive explanation asks why one outcome occurred rather than another selected alternative. The relevance of either explanation depends on the user's actual question, the selected contrast case, and the decision context [13]. A technically valid alternative may therefore remain unhelpful when it answers the wrong contrast or focuses on factors outside the stakeholder's concern.

Counterfactual quality cannot be reduced to finding the mathematically smallest change around a decision boundary. A useful account may need to satisfy validity, proximity, sparsity, plausibility, diversity, robustness, and context-sensitive actionability requirements [14]. Minimal change in a normalized feature space may correspond to an unrealistic, inaccessible, costly, or semantically incoherent change in the stakeholder's life. Feasibility constraints should therefore account for immutable attributes, causal dependencies, institutional rules, unequal burdens, and the possibility that apparently similar changes have different consequences across groups.

Causal explanations address mechanisms, interventions, or pathways rather than only modeled associations. Counterfactual language can resemble causal reasoning, but a model-relative statement that changing a feature would alter a prediction does not independently demonstrate that changing the corresponding real-world condition would cause the desired outcome [15]. Causal claims require justified assumptions, appropriate evidence, domain knowledge, and

explicit uncertainty. Without these safeguards, an explanation can convert correlational model behaviour into an unsupported narrative of cause, responsibility, or blame.

Recourse-oriented explanations extend beyond describing an alternative prediction by asking whether a person can perform a feasible action that will produce a meaningful change in the real decision process. Contrastive explanations, counterfactual examples, consequential recommendations, and effective recourse should therefore remain distinct taxonomy categories [16]. An explanation may support contestability by revealing decision boundaries or questionable factors, but contestability also requires accessible procedures, review authority, and a possibility of correction. Counterfactual explanation alone does not guarantee recourse, and causal-looking explanation alone does not guarantee causal truth or accountability.

Example-Based, Rule-Based, and Interactive Explanations

Example-based explanations communicate model behaviour through representative, similar, or contrasting cases. Case-based accounts can connect a current decision to prior instances, while prototypes summarize characteristic features of a class. Rule-based explanations instead express conditional logic, such as the conditions associated with a prediction. These forms can support concrete reasoning, but comparative evaluation indicates that explanation form may influence persuasion and advice acceptance without necessarily improving understanding or task performance [17].

Natural-language and dialogue-based explanations can translate technical outputs into ordinary or domain-specific language and allow users to ask “why,” “why not,” “what if,” and “what next.” Interactive explanation can progressively disclose detail and refine the explanatory question as a decision develops. However, conversational fluency does not ensure fidelity, and an interactive system may produce inconsistent, incomplete, or overconfident responses across dialogue turns [18].

Explanation interfaces should be designed around the workflow in which decisions occur. Participatory design with intended users can reveal which evidence, terminology, comparisons, uncertainty statements, and interaction controls are meaningful in practice [19]. Domain experts may require detailed evidence and exception handling, whereas affected individuals may need accessible reasons and challenge routes. These needs cannot be inferred reliably from a technical explanation category alone.

Adaptive and personalized explanations may modify granularity, modality, sequencing, or terminology according to expertise and interaction context. Yet user background and initial impressions can alter mental models, trust, and interpretation of AI outputs [20]. Adaptation may therefore create explanation mismatch, overload, stereotyping, or selective omission when it is poorly governed. Personalization should remain inspectable and reversible and does not guarantee understanding or responsible explanation fit. Table 2 compares counterfactual, contrastive, causal, example-based, rule-based, interactive, and adaptive explanation types according to reasoning function, stakeholder use, actionability, accountability role, and evaluation need.

Table 2. Counterfactual, Contrastive, Causal, Example-Based, Rule-Based, Interactive, and Adaptive Explanations: Reasoning Function, Actionability, Recourse, Causal Boundaries, Similarity, Rules, Dialogue, Personalization, Cognitive Risk, Evaluation Needs, and Taxonomy Roles

Explanation type	Core explanation question	Reasoning or interaction function	Likely user or task context	Potential benefit	Failure or overclaiming risk	Validation or evaluation need
Counterfactual	What would need to change for	Alternative-state reasoning	Affected person, decision	Clarifies decision boundaries	Infeasible or unstable changes	Validity, plausibility, feasibility, robustness

	another output?		maker, developer			
Contrastive	Why outcome A rather than B?	Difference-focused reasoning	Lay user, expert, auditor	Answers focused “why” questions	Irrelevant or misleading comparison	Foil relevance, fidelity, completeness
Causal	What mechanism or intervention affects the outcome?	Mechanistic reasoning	Expert, auditor, accountability actor	Supports intervention analysis	Correlation presented as causation	Identification, assumptions, uncertainty
Recourse-oriented	What feasible action may improve a future outcome?	Consequential recommendation	Adversely affected individual	Supports possible response	Action may not alter the real process	Cost, accessibility, persistence, effectiveness
Example-based	Which representative cases clarify the output?	Similarity and analogy	Lay user, expert, decision maker	Concrete comparison	Cherry-picked or biased examples	Representativeness, diversity, similarity validity
Case-based	Which prior cases resemble this case?	Precedent-style comparison	Domain and operational decisions	Supports experiential reasoning	Historical cases may be outdated or biased	Provenance, relevance, comparability
Prototype	What representative instance characterizes a class?	Category summarization	Classification tasks	Compact mental-model support	Conceals within-class variation	Coverage and subgroup representativeness
Rule-based	Which conditions led to the result?	Conditional reasoning	Expert, operator, auditor, affected person	Traceability and concise logic	Rule may oversimplify the model	Fidelity, coverage, exceptions, complexity
Natural-language	How can reasons be communicated accessibly?	Linguistic presentation	Lay user, expert, affected person	Improves accessibility	Fluent but unfaithful explanation	Factuality, fidelity, readability, uncertainty
Interactive	What follow-up information does the user need?	Iterative exploration	End user, expert, operator	Progressive disclosure and user control	Inconsistency, overload, privacy risk	Task success, coherence, cognitive load
Dialogue-based	Why, why not, what if, or	Conversational question	Decision-support	Clarifies user intent	Contradictory or	Turn-level fidelity and recovery testing

	what next?	answering	interaction		overconfident answers	
Adaptive	How should explanation content change with context?	Context-sensitive adjustment	Mixed-expertise environments	Potentially improves relevance	Stereotyping or hidden omission	Fairness, privacy, user control, longitudinal testing
Personalized	Which explanation properties suit this individual?	Individual tailoring	User-facing intelligent systems	May reduce unnecessary complexity	Preference may conflict with completeness	Preference–performance comparison
Dynamic	How should the explanation evolve over time?	Temporal updating	Monitoring and sequential decisions	Reflects changing evidence	Conflicting explanations across time	Temporal consistency and traceability
Explanation-overload boundary	Has detail exceeded processing capacity?	Cognitive-risk control	Time-pressured or non-expert users	Supports layered delivery	Reduced comprehension or verification	Cognitive-load and performance testing
Personalization boundary	Does tailoring prove responsible stakeholder fit?	Sufficiency control	All users	Prevents overclaiming	Manipulative, discriminatory, or incomplete adaptation	User override, fairness, privacy, usefulness

Stakeholder-Adaptive Explanation Requirements

Lay users and end users generally need explanations that communicate relevant reasons without unnecessary technical complexity. Accessibility may depend on familiar language, meaningful examples, restrained detail, and explicit uncertainty. End-user research indicates that explanation usefulness must be assessed in relation to the user's task rather than assumed from the availability of an explanation interface [21]. A concise local explanation may sometimes be sufficient for orientation, while a consequential decision may require layered access to supporting evidence.

Affected individuals and decision subjects may need to understand why a particular outcome occurred, whether relevant information was used correctly, and how the decision can be questioned. Preferences for explanation content and presentation can differ across users and contexts [22]. Preference, however, is not equivalent to fidelity, usefulness, or procedural adequacy. A preferred explanation may still conceal uncertainty, omit material information, or fail to provide a meaningful challenge route.

Trust should be treated as a calibration problem rather than an outcome to maximize. Perceived explainability and causability may influence user acceptance and confidence [23], but a persuasive explanation can increase confidence even when the underlying prediction or explanation is unreliable. Explanation quality therefore requires separation of perceived clarity from technical fidelity, decision usefulness, and the user's ability to recognize system limitations.

Managers, auditors, regulators, compliance actors, human overseers, and accountability actors require system-level evidence, documentation, traceability, and links between individual decisions and organizational controls. Decision makers additionally need explanations that support verification rather than passive acceptance, because reductions in overreliance depend on whether users engage with and can assess the explanation [25]. High-stakes use demands stronger validation, human control, and contestability than low-stakes orientation. **Table 3** maps stakeholder-adaptive explanation requirements across user roles, decision timing, explanation type, explanation quality, failure risks, and governance needs.

Table 3. Stakeholder-Adaptive Explanation Requirements for Explainable Intelligent Systems: Lay Users, Domain Experts, End Users, Affected Individuals, Decision Subjects, Decision Makers, Operators, Managers, Developers, Model Builders, Auditors, Regulators, Compliance Actors, Human Overseers, Decision Timing, Explanation Quality, and Stakeholder-Fit Boundaries

Stakeholder or user role	Core explanation need	Decision timing	Explanation type likely to be relevant	Required explanation quality	Failure risk	Evaluation or governance need	
Lay user	Accessible account of what happened	During or post-decision	Local, contrastive, example-based, natural-language	Clarity, relevance, uncertainty	Overload or causal misunderstanding	Comprehension and accessibility testing	User experience
Domain expert	Evidence aligned with professional reasoning	Pre-, during-, and post-decision	Global, local, rule-based, causal, interactive	Fidelity, completeness, domain relevance	Plausible but operationally irrelevant output	Domain validation and workflow testing	Expertise
End user	Understand system behaviour sufficiently for safe use	During interaction	Local, example-based, interactive, adaptive	Usefulness, accessibility, consistency	Overreliance or abandonment	Usability and reliance evaluation	User experience
Affected individual	Understand and challenge a consequential outcome	Post-decision	Local, contrastive, counterfactual, recourse-oriented	Case relevance, feasibility, procedural clarity	Illusory recourse	Accessible review and correction process	Expertise
Decision subject	Know why the system treated this case as it did	Post-decision	Local, rule-based, counterfactual	Fidelity, understandable reasons, uncertainty	Attribution interpreted as blame	Case review and challenge pathway	Expertise

Decision maker	Assess whether to accept, question, or override advice	During decision	Local, rule-based, example-based, interactive	Verification support and calibrated confidence	Automation bias or underuse	Reliance and override testing	Decision maker
Operator	Recognize anomalies and intervention needs	During operation	Global, local, dynamic, rule-based	Timeliness, stability, traceability	Alert fatigue or stale explanations	Monitoring and escalation governance	Operator
Manager	Understand aggregate performance and risk	Pre-deployment and monitoring	Global, risk summary, rule-based	Completeness, comparability, uncertainty	False assurance from summaries	Governance review and outcome monitoring	Manager
Developer	Diagnose model and data behaviour	Development and monitoring	Attribution, saliency, global, local, surrogate	Technical fidelity and reproducibility	Optimizing explanation appearance	Robustness and version testing	Developer
Model builder	Inspect learned relationships and failure regions	Development	Global, local, feature attribution, examples	Stability, coverage, diagnostic value	Misdiagnosis from unfaithful outputs	Technical benchmarking	Model builder
Auditor	Reproduce and scrutinize decisions	Pre-deployment and retrospective	Global, local samples, rules, traceability	Provenance, completeness, reproducibility	Selective disclosure	Independent access and audit authority	Auditor
Regulator, where supported	Assess whether claims and controls are substantiated	Pre-deployment or investigation	Global, accountability-oriented, causal where valid	Scope clarity, evidence quality, uncertainty	Explanation treated as legal sufficiency	Independent technical assessment	Regulator, where supported
Compliance actor, where supported	Connect system behaviour with internal controls	Pre-deployment and monitoring	Global, documentation-linked, rule-based	Traceability and coverage	Checklist compliance without validation	Control testing and escalation	Compliance actor, where supported
Human overseer	Know when and how to intervene	During decision or operation	Local, dynamic, uncertainty-aware, interactive	Timeliness and actionability	Nominal oversight without control	Override usability and authority testing	Human overseer

Accountability actor	Determine roles, evidence, and responsibility	Post-event	Traceability, global/local evidence, causal where justified	Provenance and role clarity	Responsibility displaced onto the interface	Responsibility mapping and record retention	Accountability evidence
High-stakes decision user	Verify consequential advice under uncertainty	Before and during decision	Interpretable, local, uncertainty-aware, causal where validated	Robustness, domain validity, contestability	Serious harm from false reassurance	Independent validation and monitoring	Risk evidence
Low-stakes decision user	Obtain proportionate orientation	During decision	Concise local, example-based, natural-language	Relevance and efficiency	Disproportionate cognitive burden	Proportionate usability evaluation	Local evidence
Stakeholder-fit boundary	Avoid assigning one explanation solely by role	All moments	Multiple candidate types	Demonstrated contextual fit	Stereotyping and explanation mismatch	User-centered validation and choice	Stakeholder evidence

Proposed Explainability Taxonomy

The proposed taxonomy begins with explanation purpose rather than method availability. It distinguishes system-scope explanations for general model behaviour from instance-scope explanations for particular outputs. Selection then proceeds through stakeholder role, decision timing, consequence, required cognitive accessibility, and the evidence needed to evaluate the explanation. Because explanation quality contains distinct properties, no single evaluation measure can represent fidelity, completeness, stability, plausibility, usefulness, and accessibility simultaneously [26].

A second axis differentiates reasoning functions. Actionability-oriented explanations describe possible changes; reasoning-oriented explanations answer contrastive questions; causality-oriented explanations address mechanisms under justified assumptions; similarity-oriented explanations use cases or prototypes; and rule-oriented explanations express conditional logic. Interaction-oriented and adaptation-oriented explanations describe how explanatory content is explored or tailored. Human-centered evaluation remains necessary because explanation effects depend on the task, interface, user, and study context [27].

A third axis concerns stakeholder and risk orientation. Developer-facing explanations prioritize diagnosis and reproducibility, user-facing explanations prioritize accessible decision support, affected-person explanations prioritize reasons and challenge, and auditor-facing explanations prioritize traceability and reproducibility. Risk-sensitive explanations require stronger evidence where consequences are substantial. In high-stakes domains, technical truthfulness, human plausibility, relevance, and understandability should be assessed separately rather than collapsed into apparent usefulness [28].

A fourth axis covers accountability, contestability, and oversight. Accountability-oriented explanations connect evidence with responsible actors and controls; contestability-oriented explanations support informed challenge; and oversight-oriented explanations support monitoring, escalation, and human intervention. Explainability may contribute to trustworthy AI, but it is only one component among broader requirements concerning robustness, fairness, governance, safety, and

human control [29]. The taxonomy therefore organizes explanation requirements without claiming that explanation establishes trustworthy operation.

The categories are connected rather than mutually exclusive. A local counterfactual can be instance-scope, actionability-oriented, affected-person-facing, and contestability-oriented at the same time. More explanation is not always better because comprehensibility, completeness, confidentiality, confidence, and responsible conduct can create competing design pressures [30]. Trust also depends on user, task, system, and relational conditions beyond explanation type [31]. Table 4 presents the proposed stakeholder-oriented explainability taxonomy linking explanation categories, explanation types, stakeholder fit, decision timing, evaluation requirements, failure risks, and boundary statements.

Table 4. Proposed Stakeholder-Oriented Explainability Taxonomy: System-Scope, Instance-Scope, Actionability-Oriented, Reasoning-Oriented, Causality-Oriented, Similarity-Oriented, Rule-Oriented, Interaction-Oriented, Adaptation-Oriented, Accountability-Oriented, Contestability-Oriented, Oversight-Oriented, Developer-Facing, User-Facing, Affected-Person, Auditor-Facing, and Risk-Sensitive Explanations

Taxonomy category	Core purpose	Explanation types included	Primary stakeholder fit	Decision timing	Evaluation requirement	Failure risk	
System-scope	Describe general model behaviour	Global summaries, transparent models, surrogates, rules	Developer, auditor, manager, expert	Development, pre-deployment, monitoring	Fidelity, coverage, stability	Aggregation hides local failures	Do not
Instance-scope	Explain a particular output	Local attribution, saliency, local rule, example	Affected person, decision maker, operator	During or post-decision	Local fidelity, relevance, stability	Case explanation generalized system-wide	Global
Actionability-oriented	Identify possible next steps	Counterfactual, actionable, recourse-oriented	Affected person, operator	Post-decision and planning	Feasibility, cost, robustness	Impossible or ineffective change	
Reasoning-oriented	Explain why one outcome rather than another	Contrastive, rule-based, natural-language	Lay user, expert, auditor	During or post-decision	Foil relevance, fidelity, completeness	Wrong explanatory question answered	
Causality-oriented	Describe mechanisms or interventions	Causal pathways, structural counterfactuals	Expert, auditor, accountability actor	Planning and retrospective analysis	Identification, assumptions, uncertainty	Correlation presented as causation	Causation
Similarity-oriented	Explain through comparable instances	Examples, cases, prototypes, neighbours	Lay user, expert, decision maker	During or post-decision	Representativeness and similarity validity	Biased or cherry-picked examples	Similarity

Rule-oriented	Communicate conditional logic	Intrinsic rules, extracted rules, local rules	Expert, operator, auditor, affected person	Development, decision, audit	Fidelity, coverage, complexity	False impression of determinism	Ex
Interaction-oriented	Support questioning and exploration	Interactive, conversational, dialogue-based	End user, expert, operator	During or post-decision	Coherence, turn fidelity, task success	Inconsistency or overload	ur
Adaptation-oriented	Adjust content, detail, or modality	Adaptive, personalized, layered, dynamic	Mixed-expertise users	All moments	Fairness, privacy, user control	Stereotyping or hidden omission	Pe
Accountability-oriented	Connect decisions with evidence and responsibility	Traceability, documentation, causal where justified	Auditor, manager, accountability actor	Pre-deployment and retrospective	Provenance, completeness, reproducibility	Explanation substitutes for governance	a
Contestability-oriented	Enable informed challenge and correction	Local, contrastive, counterfactual, evidence-linked	Affected person, reviewer	Post-decision	Accessibility and review effectiveness	Information without remedy	c
Oversight-oriented	Support monitoring and intervention	Dynamic, uncertainty-aware, global/local diagnostics	Operator, overseer, manager	During operation	Timeliness and override usability	Nominal oversight	H
Developer-facing	Diagnose model and data behaviour	Attribution, saliency, global, local, surrogate	Developer, model builder	Development and monitoring	Technical correctness and reproducibility	Confirmation bias	in
User-facing	Support informed system use	Local, example-based, natural-language, interactive	End user, decision maker	During or post-decision	Accessibility, usefulness, uncertainty	Persuasion without verification	S
Affected-person	Support understanding, challenge, and response	Local, contrastive, counterfactual, recourse-oriented	Decision subject, affected individual	Post-decision	Case relevance and procedural accessibility	Illusory recourse	er
Auditor-facing	Support independent scrutiny	Global, local samples,	Auditor, compliance actor	Pre-deployment	Access, reproducibility, versioning	Selective evidence	

		traceability, documentation		and retrospective			a
Risk-sensitive	Scale evidence to consequence	Interpretable, uncertainty- aware, causal where valid	High-stakes user and overseer	Pre-, during-, and post- decision	Domain validity, robustness, monitoring	Persuasive explanation masks risk	I do
Explanation- type sufficiency boundary	Prevent universal method claims	All types	All stakeholders	All moments	Cross-method contextual evaluation	Method- centric overclaiming	in tru
Stakeholder-fit boundary	Prevent deterministic role matching	All types	All stakeholders	All moments	Empirical fit testing and user choice	Role stereotyping	F no
Deployment- readiness boundary	Separate taxonomy from operational approval	All types	Developer, manager, auditor, operator	Pre- deployment	System, human, domain, and governance evidence	Conceptual classification treated as certification	TI do c

Figure 1 presents the proposed stakeholder-oriented explainability taxonomy, linking explanation types, stakeholder roles, decision timing, explanation purpose, evaluation requirements, failure risks, and boundary conditions.

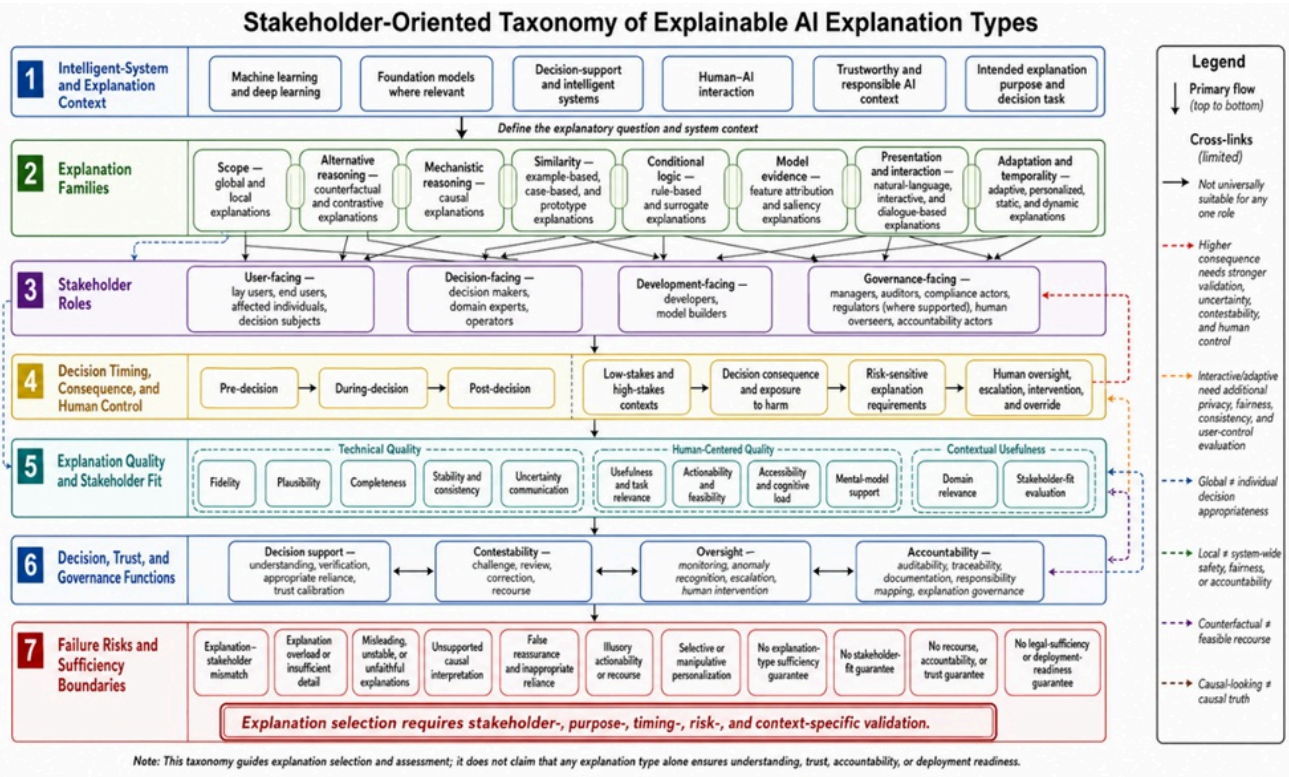


Figure 1. Stakeholder-Oriented Taxonomy of Explainable AI Explanation Types

Alt-text description:

A conceptual taxonomy showing explanation types connected to stakeholder roles, decision timing, explanation purposes, quality criteria, evaluation needs, failure risks, accountability, contestability, and boundaries against explanation sufficiency.

Evaluation and Design Implications

Explanation-type selection should begin by identifying the explanatory question, stakeholder, decision moment, consequence, and intended use. Designers should then compare candidate explanation families against purpose-relative quality requirements. Fidelity, stability, completeness, plausibility, accessibility, and usefulness should be treated as distinct properties rather than compressed into a universal explanation score [26]. Technical evaluation should also disclose the explanation's valid scope and assumptions.

Stakeholder-fit and decision-timing evaluation require user-centered testing in realistic workflows. Cognitive-accessibility testing should examine comprehension, error detection, cognitive load, and the ability to identify uncertainty. Explanation-usefulness testing should assess whether the account supports the intended task rather than merely increasing satisfaction. Human-centered studies remain necessary because explanation outcomes vary across users, domains, interfaces, and decision conditions [27].

Trust-calibration evaluation should assess whether confidence and reliance track system capability. Reliance monitoring can examine when users accept, reject, verify, or override AI advice, especially when the system is wrong. Actionability assessment should test feasibility, accessibility, cost, and robustness, while recourse evaluation should determine whether proposed actions can affect the real institutional process. Contestability further requires an accessible review route, authority to reconsider the decision, and a mechanism for correction.

Accountability mapping should connect explanations with data provenance, system versions, responsible actors, review authority, documentation, escalation, and human control. Uncertainty communication and explanation-failure detection should cover instability, inconsistency, causal ambiguity, and misleading fluency. Adaptive-explanation governance should include privacy, fairness, user inspection, and override controls because contextualization can create both benefits and adverse effects [30]. These are conceptual research and design implications, not legal, compliance, certification, procurement, platform-selection, model-selection, or deployment instructions.

Conclusion

This article proposes a stakeholder-oriented explainability taxonomy linking global, local, counterfactual, contrastive, causal, example-based, rule-based, interactive, and adaptive explanations with stakeholder roles, decision timing, explanatory purpose, quality requirements, evaluation needs, contestability, accountability, and explicit boundary conditions. Its central premise is that explanation selection should begin with the stakeholder's question, responsibility, expertise, decision context, cognitive needs, potential for action, reliance risk, and accountability requirements rather than with method availability alone. The taxonomy can organize design choices and clarify evaluation responsibilities, but it does not independently prove understanding, fairness, causal truth, recourse, trustworthiness, legal sufficiency, accountability, or deployment readiness. Explanation selection therefore remains a context-sensitive design and evaluation problem requiring technical validation, user-centered assessment, domain evidence, uncertainty communication, governance mechanisms, and meaningful human control.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Dec 2024

Revised: 25 Jan 2025

Accepted: 24 Mar 2025

Published online: 15 September 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>.
- 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>.
- 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>.
- 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>.
- 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>.

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*. 2019;116(44):22071-80.
<https://doi.org/10.1073/pnas.1900654116>.

Cabitza F, Campagner A, Malgieri G, Natali C, Schneeberger D, Stöger K, et al. Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI. *Expert Syst Appl*. 2023;213:118888.
<https://doi.org/10.1016/j.eswa.2022.118888>.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
<https://doi.org/10.1038/s42256-019-0138-9>.

Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *Artif Intell*. 2021;298:103502.
<https://doi.org/10.1016/j.artint.2021.103502>.

Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst*. 2017;28(11):2660-73.
<https://doi.org/10.1109/TNNLS.2016.2599820>.

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>.

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;9:11974-2001.
<https://doi.org/10.1109/ACCESS.2021.3051315>.

Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min Knowl Discov*. 2024;38(5):2770-824.
<https://doi.org/10.1007/s10618-022-00831-6>.

Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf Fusion*. 2022;81:59-83.
<https://doi.org/10.1016/j.inffus.2021.11.003>.

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>.

van der Waa J, Nieuwburg E, Cremers A, Neerinx M. Evaluating XAI: a comparison of rule-based and example-based explanations. *Artif Intell*. 2021;291:103404.
<https://doi.org/10.1016/j.artint.2020.103404>.

Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artif Intell*. 2021;299:103525.
<https://doi.org/10.1016/j.artint.2021.103525>.

Panigutti C, Beretta A, Fadda D, Giannotti F, Pedreschi D, Perotti A, et al. Co-design of human-centered, explainable AI for clinical decision support. *ACM Trans Interact Intell Syst*. 2023;13(4):1-35.
<https://doi.org/10.1145/3587271>.

Nourani M, Roy C, Block JE, Honeycutt DR, Rahman T, Ragan ED, et al. On the importance of user backgrounds and impressions: lessons learned from interactive AI applications. *ACM Trans Interact Intell Syst*. 2022;12(4):1-29.
<https://doi.org/10.1145/3531066>.

Laato S, Tiainen M, Najmul Islam AKM, Mäntymäki M. How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Res*. 2022;32(7):1-31.
<https://doi.org/10.1108/INTR-08-2021-0600>.

Kim D, Song Y, Kim S, Lee S, Wu Y, Shin J, et al. How should the results of artificial intelligence be explained to users? Research on consumer preferences in user-centered explainable artificial intelligence. *Technol Forecast Soc Change*. 2023;188:122343.
<https://doi.org/10.1016/j.techfore.2023.122343>.

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>.

Dikmen M, Burns C. The effects of domain knowledge on trust in explainable AI and task performance: a case of peer-to-peer lending. *Int J Hum Comput Stud*. 2022;162:102792.
<https://doi.org/10.1016/j.ijhcs.2022.102792>.

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>.

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>.

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>.

Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal*. 2023;84:102684.
<https://doi.org/10.1016/j.media.2022.102684>.

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>.

Abedin B. Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective. *Internet Res*. 2022;32(2):425-53.
<https://doi.org/10.1108/INTR-05-2020-0300>.

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>.