

REVIEW

Open access

Why Explainable AI Is Adopted, Ignored, or Misused—A Realist Review of Contexts, Trust Mechanisms, Human Responses, and Decision Outcomes

Peter Novak^{1*}, Leila Haddad², Johan Nilsson³

Abstract

Explainable artificial intelligence is often presented as a means of making intelligent systems more understandable, trustworthy, and usable, yet explanation availability does not reliably produce adoption, appropriate reliance, accountability, or improved decisions. This realist review examines why explainable AI is adopted, ignored, rejected, discontinued, overused, underused, or misused across organizational, professional, technical, and decision settings. It applies context–mechanism–outcome reasoning to interpret how explanation design interacts with workflow integration, user and domain expertise, decision risk, organizational incentives, governance arrangements, and human-oversight conditions. The synthesis focuses on mechanisms including perceived transparency, cognitive accessibility, mental-model formation, perceived usefulness, trust calibration, accountability confidence, contestability awareness, algorithm aversion, algorithm appreciation, automation bias, and explanation overload. These mechanisms may support selective reliance and meaningful challenge, but they may also generate false reassurance, superficial compliance, defensive rejection, or uncritical dependence. The review distinguishes adoption from acceptance, trust from trustworthiness, reliance from appropriate reliance, and explanation usefulness from decision quality. It identifies plausible configurations in which professional expertise, uncertainty communication, traceability, feedback, and contestability support careful use, while poor workflow fit, limited accessibility, weak governance, and automation pressure suppress scrutiny or encourage misuse. The main contribution is a bounded realist synthesis and design agenda for evaluating explainable AI through context-sensitive human responses and decision outcomes rather than explanation availability alone. The configurations are explanatory propositions requiring human-centered, organizational, and decision-context validation; they are not universal causal laws, implementation protocols, or evidence of deployment readiness.

Keywords Explainable AI, Trust calibration, Realist review, AI adoption, Human reliance, Automation bias

*Correspondence:

Peter Novak

peter.novak@cuni.cz

¹ Department of AI Adoption and Governance, Charles University, Prague, Czech Republic

² Department of Information Systems, American University of Beirut, Beirut, Lebanon

³ Department of Informatics and Systems Science, Uppsala University, Uppsala, Sweden

Introduction

Explainable artificial intelligence has become central to trustworthy intelligent systems because users, organizations, and affected persons may need reasons for predictions, recommendations, or automated decisions. However, explanation is a communicative and selective process shaped by the recipient, question, contrast, and social setting rather than a simple transfer of technical information [1].

The XAI field contains heterogeneous objectives, including interpretability, transparency, debugging, trust support, accountability, and responsible system design. These objectives are not interchangeable, and an explanation designed for developers may not satisfy the informational or procedural needs of decision makers, operators, auditors, managers, or affected persons [2].

Technical methods also produce different explanatory objects. Feature importance, local rules, examples, counterfactuals, and global model summaries answer different questions and impose different assumptions. Consequently, explanation quality cannot be assessed independently of the decision problem, intended user, model characteristics, and practical purpose for which an explanation is requested [3].

The expansion of XAI research therefore does not establish that explanations are adopted, understood, or used appropriately. This review examines adoption, rejection, non-use, disuse, overuse, underuse, and misuse through realist context–mechanism–outcome reasoning. Its contribution is to explain variable human responses while maintaining explicit no-adoption, no-decision-quality, and no-deployment-readiness boundaries [4].

Explainable AI Adoption as a Context-Dependent Process

Explainable-AI adoption is not a single transition from opacity to acceptance. Explainability can activate confidence, comprehensibility, and perceived usefulness while also generating confidentiality concerns, cognitive burden, strategic disclosure, or false assurance. These opposing effects make adoption contingent on organizational conditions, stakeholder goals, explanation properties, and decision consequences [5].

Users also approach AI-supported decisions with different levels of epistemic uncertainty. When users are unsure about their own judgment, an explanation may increase advice acceptance; when they possess stronger evidence or perceive inconsistency, the same explanation may be discounted. Adoption therefore reflects interaction between explanation design and the user's perceived knowledge position [6].

Stakeholders differ in what they expect explanations to accomplish. Developers may seek debugging information, managers may seek operational justification, professionals may need decision-relevant evidence, auditors may require traceability, and affected persons may need understandable reasons and recourse. An explanation can consequently be useful to one stakeholder while remaining inaccessible or irrelevant to another [7].

Adoption must also be distinguished from acceptance, intention, and beneficial use. Narrow local explanations may encourage initial engagement yet fail to support deeper questions about system goals, causal assumptions, uncertainty, or limitations. Context-sensitive adoption analysis must therefore examine cognitive accessibility, continued use, decision fit, misuse risks, and whether explanation depth supports the required human–AI interaction [8].

Realist Review Scope and Context–Mechanism–Outcome Logic

The review asks: in which contexts is XAI adopted, ignored, rejected, or misused; which mechanisms connect explanation design with trust, understanding, usefulness, reliance, accountability, and contestability; how do aversion, appreciation, automation bias, overreliance, and underreliance shape use; what outcomes follow; which CMO configurations explain variation; and which priorities may support safer, more useful implementation [9].

A context is a condition that shapes whether a mechanism becomes active, including organizational readiness, professional role, technical quality, workflow, expertise, incentives, governance, time pressure, and decision risk. A mechanism concerns how actors interpret and respond to available resources. Outcomes include adoption, non-use,

rejection, reliance, misuse, learning, challenge, accountability, and decision consequences. Explainability is therefore not treated as one stable intervention [10].

Evidence was selected from peer-reviewed Q1 journal articles with DOI information and relevance to XAI, human–AI interaction, organizational adoption, trust, reliance, accountability, recourse, or decision use. Extraction focused on reported contexts, mechanisms, outcomes, human responses, implementation conditions, and limitations. Technical demonstrations without usable human, organizational, or decision evidence were not treated as sufficient support for adoption claims [11].

The synthesis compares recurring evidence patterns rather than calculating pooled adoption or trust effects. It distinguishes directly reported findings from proposed realist interpretations and treats every configuration as conditional. Evidence that human–AI combinations vary by task reinforces the boundary that system use, explanation provision, or human oversight cannot independently establish complementary performance or improved decision quality [12].

Table 1 summarises the realist review logic used to map explainable AI adoption through context, mechanism, outcome, evidence extraction, and synthesis categories.

Table 1. Realist Review Logic for Explainable AI Adoption: Review Questions, Context–Mechanism–Outcome Reasoning, Inclusion Criteria, Exclusion Criteria, Evidence Extraction Categories, Context Mapping, Mechanism Mapping, Outcome Mapping, Citation Planning, and Review Boundaries

<i>Review component</i>	<i>Core realist question or purpose</i>	<i>Inclusion logic</i>	<i>Exclusion logic</i>	<i>Evidence mapped</i>	<i>CMO dimension addressed</i>	<i>Bias or limitation risk</i>	<i>Realist review r</i>
<i>Review questions</i>	Why does XAI use vary across settings?	Evidence explaining contexts, responses, or outcomes	Purely descriptive method papers	Adoption, rejection, misuse, reliance	Cross-cutting	Questions may privilege known mechanisms	Direct explanat synthes
<i>Evidence eligibility</i>	Is the source suitable for planned claims?	Peer-reviewed Q1 journal article with DOI	Conference, preprint, report, or non-peer-reviewed source	Source type and claim relevance	Cross-cutting	Venue screening does not guarantee evidence quality	Establish minimu eligibilit
<i>XAI relevance</i>	Does the article concern explanation or AI-supported use?	XAI, opacity, trust, advice, accountability, or recourse relevance	Generic adoption without AI relevance	System and explanation characteristics	Context	Broad adjacent literature may be remote	Define phenome boundar
<i>Context mapping</i>	Under what conditions does a response occur?	Organizational, professional, technical, workflow, expertise, or risk evidence	Context-free generalization	Setting, role, task, incentives, governance	Context	Some studies report context incompletely	Identifie enabling suppress conditio

<i>Mechanism mapping</i>	How do actors interpret available explanations?	Trust, usefulness, accessibility, agency, burden, aversion, or accountability evidence	Unsupported psychological speculation	Human reasoning and response	Mechanism	Mechanisms may be inferred rather than measured	Explains v resource produc differer respons
<i>Outcome mapping</i>	What follows from explanation use or non-use?	Behavioral, decision, implementation, or accountability outcomes	Technical accuracy without human-use relevance	Adoption, reliance, rejection, misuse, performance	Outcome	Intentions may be mistaken for behavior	Distinguish attitudes f consequer
<i>CMO configuration</i>	Which context activates which mechanism and outcome?	Plausible evidence-linked configurations	Universal causal claims	Recurrent cross-study patterns	C–M–O	Associations may be overinterpreted	Produce bounde realist explanati
<i>Evidence extraction</i>	What information supports each interpretation?	Reported aims, settings, mechanisms, outcomes, and limitations	Invented or unreported details	Structured article-level evidence	Cross-cutting	Selective extraction	Maintain traceabi
<i>Citation planning</i>	Does each citation support the exact sentence?	Direct or carefully bounded conceptual support	Random or convenience citation	Claim–source alignment	Cross-cutting	Citation clustering may conceal weak fit	Preserv Vancouv sequenc and evidenc fit
<i>Review boundaries</i>	What cannot be inferred?	Explicit uncertainty and generalization limits	Readiness, certification, or guaranteed-benefit claims	Evidence gaps and transfer limits	Cross-cutting	Readers may treat configurations as rules	Protect realist-rev integrit

Figure 1 illustrates the realist review context–mechanism–outcome logic used to explain why explainable AI is adopted, ignored, rejected, or misused across decision settings.

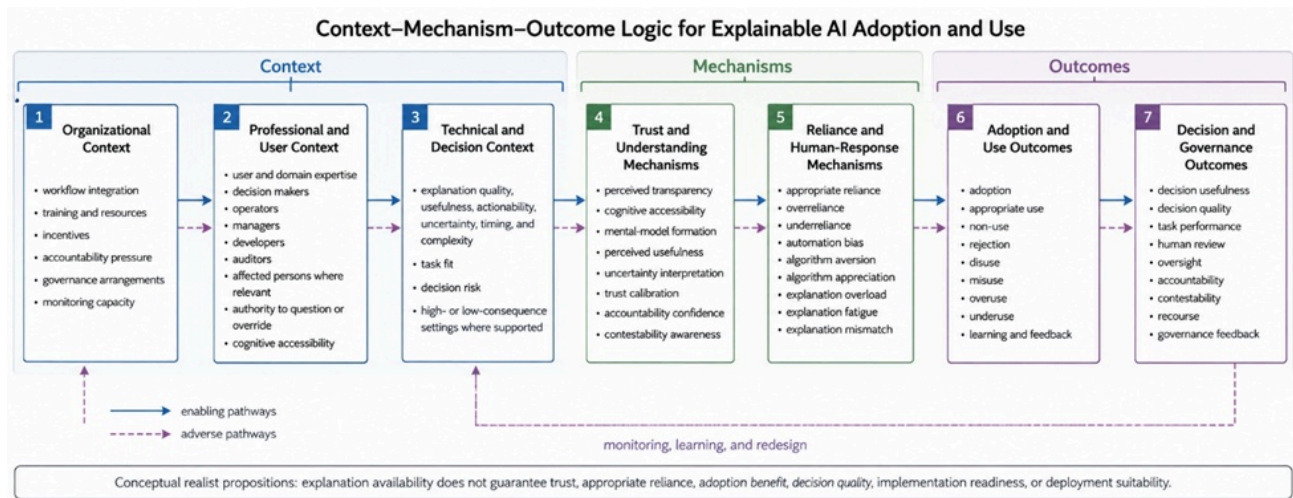


Figure 1. Context–Mechanism–Outcome Logic for Explainable AI Adoption and Use

Alt-text description:

A realist review diagram showing contexts feeding into mechanisms such as trust calibration, cognitive accessibility, perceived usefulness, accountability, and automation bias, which then lead to adoption, rejection, misuse, reliance, and decision outcomes.

Organizational, Professional, and Decision Contexts

Organizational adoption depends on whether an institution possesses suitable data, skills, leadership commitment, governance, resources, and a defined purpose for AI use. Readiness is therefore purpose-specific rather than a general organizational attribute. Explanations introduced without compatible capabilities, monitoring, or responsibility structures may produce symbolic adoption rather than sustained and meaningful use [13].

Professional context shapes how AI output is interrogated. Domain experts may compare recommendations with clinical, operational, or experiential knowledge, but expertise does not automatically generate calibrated reliance. Professional engagement also depends on workflow, ambiguity-management practices, authority to challenge outputs, and whether the organization permits users to investigate disagreements rather than merely accept recommendations [14].

Decision context determines the evidentiary burden placed on explanations. High-stakes and risk-sensitive decisions require stronger validation, uncertainty communication, and review than reversible low-stakes tasks. Medical-XAI scholarship illustrates why technical interpretability must be connected with domain knowledge, education, accountability, and human evaluation before explanations can be considered suitable for professional decision support [15].

Human oversight must likewise be assessed as a contextual condition rather than a binary safeguard. Decision makers, operators, managers, developers, and auditors may have different information, time, authority, and incentives. In high-pressure settings, nominal review may become rubber-stamping, while persuasive explanations can conceal system weaknesses. Current explanations cannot independently establish clinical trustworthiness or implementation readiness [16].

User and domain expertise also affect accessibility. Lay users and affected persons may need plain-language reasons, consequences, and challenge routes, whereas expert users may require technical evidence, uncertainty, and traceability. Evidence from stakeholder-tailored clinical interfaces shows that clinicians and patients can require different explanatory content and presentation, limiting context-generalization across roles and decision environments [17].

Trust, Understanding, and Reliance Mechanisms

Trust is an attitude concerning expected system behavior, whereas trustworthiness concerns whether that confidence is warranted by actual properties and performance. Reliance is behavioral and may diverge from stated trust. Organizational representation, task characteristics, system embodiment, prior experience, and user expectations therefore shape trust without necessarily producing appropriate use [18].

Perceived transparency and causability can influence whether users feel able to understand an AI recommendation. Such perceptions may increase acceptance by making system behavior appear more intelligible and useful, but perceived explainability is not equivalent to faithful explanation. Trust formed through a persuasive interface may remain disconnected from model validity or decision suitability [19].

Trust calibration depends on users receiving credible information about both strengths and limitations. Explanation effects may differ according to model accuracy and whether explanatory content appears coherent or nonsensical. Explanations should therefore be evaluated alongside performance evidence, because intelligible presentation can otherwise increase confidence in systems whose recommendations should be questioned [20].

Cognitive accessibility and explanation usefulness are similarly insufficient when they produce understanding without calibration. In clinical decision support, understandable explanation classes may increase reliance while failing to help users distinguish correct from incorrect advice. Explanation design should therefore support uncertainty interpretation, verification, and recognition of failure conditions rather than merely increasing confidence [21].

Appropriate reliance occurs when behavior reflects system competence, task fit, and available contrary evidence. In a defined high-risk task, explanations may support trust and performance, but such effects remain dependent on the system, user, and decision setting [22]. Explanation fatigue, overload, algorithm aversion, algorithm appreciation, and automation bias should consequently be treated as conditional mechanisms, with no trust-guarantee or reliance-benefit guarantee.

Adoption, Rejection, Misuse, and Decision Outcomes

Rejection and underuse may arise when users observe algorithmic errors and feel unable to influence the resulting judgment. Providing limited opportunities to adjust an imperfect algorithm can restore agency and increase willingness to use it. This mechanism may reduce algorithm aversion, but greater adoption does not independently establish that the resulting decisions are more accurate or appropriate [23].

The opposite response is also possible. Users may assign greater weight to algorithmic advice than to human advice when they perceive the system as objective, consistent, or comparatively competent. Algorithm appreciation can support useful adoption, but it can also encourage overuse when perceived authority exceeds demonstrated capability or when users neglect contextual evidence [24].

Use patterns vary with the perceived relationship between an algorithm and its task. People may be more receptive when a task appears objective and rule-governed and more resistant when it appears subjective or personally sensitive. Adoption, rejection, and reliance should therefore be interpreted against perceived task fit rather than treated as stable attitudes toward AI [25].

Algorithm aversion is produced by interacting user, task, system, and environmental conditions. Non-use may reflect error sensitivity, loss of agency, professional judgment, inadequate explanation, or rational concern about consequences. It should not automatically be classified as poor understanding, just as high use should not automatically be treated as successful implementation [26].

Table 2 maps organizational, professional, technical, and decision contexts to trust, understanding, reliance, adoption, rejection, misuse, and decision outcomes in explainable AI use.

Table 2. Contexts, Mechanisms, and Outcomes in Explainable AI Use: Organizational Context, Professional Context, Decision Context, Workflow Integration, User Expertise, Trust Calibration, Cognitive Accessibility, Reliance, Automation Bias, Algorithm Aversion, Adoption, Rejection, Misuse, and Decision Outcomes

Context, mechanism, or outcome component	Core realist question	Relevant condition or trigger	Human response or mechanism pathway	Potential use pattern or outcome	Failure or misuse risk	Monitoring or evaluation need	
Organizational context	Can the institution support meaningful XAI use?	Skills, resources, governance, data, leadership	Feasibility and organizational confidence	Adoption, piloting, or non-use	Symbolic adoption	Readiness and longitudinal-use assessment	F
Professional context	How do role obligations affect interpretation?	Expertise, liability, norms, authority	Critical interrogation or defensive avoidance	Selective use or rejection	Uncritical deference or identity-based dismissal	Role-specific field evaluation	Ex
Decision context	What consequences and reversibility apply?	High, low, or risk-sensitive stakes	Caution, urgency, or perceived usefulness	Reliance, escalation, or deferral	Inappropriate transfer across tasks	Decision-specific outcome assessment	Fir be
Workflow integration	Is explanation available at the right point and cost?	Timing, workload, interruption, interface fit	Reduced burden or disruption	Routine use or disuse	Skimming and bypassing	In-situ workflow observation	qu
User expertise	Can the intended user interpret the explanation?	AI literacy, domain knowledge, language, experience	Cognitive accessibility or mismatch	Understanding, non-use, or false reassurance	Jargon or oversimplification	Expertise-stratified comprehension testing	c c
Trust calibration	Does confidence track system capability?	Performance evidence and visible limitations	Confidence increases or decreases appropriately	Selective reliance	Blind trust or excessive distrust	Compare trust with accuracy awareness and behavior	M nc

<i>Mental-model formation</i>	Can users anticipate strengths, limits, and errors?	Repeated interaction, feedback, contrastive information	Understanding and expectation updating	Verification and learning	Confident but inaccurate mental model	Longitudinal prediction and comprehension tests	co do
<i>Cognitive accessibility</i>	Is explanation effort manageable?	Complexity, time pressure, disability, expertise	Comprehension or overload	Use, superficial use, or non-use	False simplification	Workload and accessibility assessment	R ne un
<i>Explanation actionability</i>	Can users respond to the explanation?	Feasible recourse or decision alternatives	Challenge, correction, or adaptation	Contestability and remediation	Symbolic explanation without practical route	Recourse accessibility and outcome tracking	Re d
<i>Appropriate reliance</i>	Does behavior reflect system competence and task fit?	Calibrated trust and verification opportunity	Selective advice acceptance	Potentially improved decisions	Reliance measured only by frequency	Behavioral and outcome-linked evaluation	R be c
<i>Overreliance</i>	Is AI followed despite contrary evidence?	Automation pressure and authoritative presentation	Deference and reduced checking	Overuse or misuse	Costly or harmful decisions	Contradiction tests and advice-acceptance logs	se
<i>Underreliance</i>	Is competent assistance discounted?	Prior error, weak fit, identity conflict	Aversion or trust erosion	Underuse or rejection	Missed complementarity	Compare behavior with validated system competence	R
<i>Automation bias</i>	Does AI become the default anchor?	Time pressure, workload, repeated automation	Reduced independent search	Commission or omission errors	Rubber-stamping	Review time, override, and escalation monitoring	pr ne
<i>Algorithm aversion</i>	Does visible imperfection prompt rejection?	Errors, low control, subjective task	Loss of confidence or agency	Rejection and underuse	Dismissing useful assistance	Reason-for-rejection analysis	
<i>Algorithm appreciation</i>	Is algorithmic advice given excessive authority?	Objectivity cues and low user confidence	Preferential advice weighting	Adoption or overuse	Overconfidence in automation	Compare source weighting with competence	A u

<i>Adoption</i>	Is XAI incorporated into practice?	Readiness, usefulness, workflow fit	Acceptance and repeated engagement	Initial or sustained use	Adoption without scrutiny	Distinguish trial from routine institutional use	A
<i>Non-use and disuse</i>	Why is available XAI ignored or abandoned?	Burden, poor fit, errors, weak support	Avoidance or trust erosion	Bypass or discontinuation	Lost benefit or avoided harm	Longitudinal use and qualitative evidence	N
<i>Misuse and overuse</i>	Is AI used beyond suitable scope or without safeguards?	Weak governance and convenience incentives	False reassurance or excessive delegation	Unsupported application	Harm, deskilling, responsibility diffusion	Incident and scope-violation monitoring	Hi
<i>Decision and accountability outcomes</i>	Does XAI improve decisions or answerability?	Valid system, appropriate allocation, governance	Verification, traceability, challenge	Better, unchanged, or worse outcomes	Explanation used as liability shield	Comparative performance and responsibility assessment	M im

Realist Synthesis of Explainable AI Use

In contestable decisions, explanations become more meaningful when they support feasible action, correction, or appeal. Algorithmic recourse research indicates that contrastive reasons and consequential recommendations require contextual and causal validity; superficially actionable advice may otherwise direct affected persons toward changes that are ineffective, inaccessible, or unrelated to the actual decision process [28].

Transparency alone provides no universal accountability pathway. Making selected system elements visible may produce an impression of openness without giving users meaningful knowledge, institutional authority, or control. In weak-governance settings, perceived transparency may therefore activate false reassurance or symbolic compliance rather than scrutiny and correction [29].

High-accountability contexts can support more careful use when explanations are linked to traceability, assigned responsibility, and genuine review authority. Transparency may contribute to accountability only when institutions specify who owns the system, who reviews its outputs, who can intervene, and how adverse decisions are challenged. Without these arrangements, responsibility may remain diffuse [30].

Perceived fairness, accountability, and transparency can influence trust and willingness to use algorithmic services [31]. Yet such perceptions do not prove that a system is actually fair, accountable, or transparent. Realist interpretation must therefore distinguish governance-related confidence from institutional performance and examine whether trust activates appropriate scrutiny, passive acceptance, defensive use, or rejection.

Affected persons may reject high-stakes AI when they believe it cannot recognize individual circumstances. Resistance to medical AI can be explained partly by perceived uniqueness neglect rather than a general inability to understand technology [32]. Explanation form also matters: rule-based and example-based explanations may influence understanding, persuasion, and performance differently, requiring these outcomes to be evaluated separately [33].

Table 3 presents the realist synthesis of explainable AI use, including context–mechanism–outcome configurations, interpretation, design implications, residual risks, and implementation priorities.

Contestability, Human Oversight, Decision Outcomes, Residual Risks, and Design Agenda

<i>CMO configuration or synthesis theme</i>	<i>Context involved</i>	<i>Mechanism involved</i>	<i>Outcome involved</i>	<i>Evidence pattern</i>	<i>Design or implementation implication</i>	<i>Residual risk</i>	
<i>Accountability-supported careful use</i>	Clear responsibility, review authority, professional accountability	Explanation usefulness and traceability	Verification and selective use	Professionals engage more critically when they can interrogate and justify decisions	Connect explanations to records, escalation, and assigned ownership	Defensive documentation or blame shifting	E a n
<i>Accessibility failure</i>	Low expertise, dense presentation, time pressure	Cognitive burden and explanation mismatch	Superficial use or non-use	Stakeholders require different explanatory content and depth	Tailor language, format, timing, and detail	Oversimplification may conceal uncertainty	E e a
<i>Automation-pressure overreliance</i>	High workload, authoritative AI, weak review	Trust without verification and automation bias	Overuse or misuse	Users may follow advice despite conflicting evidence	Add contradiction prompts and risk-based escalation	Safeguards may become routine clicks	o p
<i>Weak-trust rejection</i>	Visible errors, low control, poor task fit	Algorithm aversion	Rejection or underuse	Control and task framing alter willingness to use algorithms	Provide bounded agency and credible performance evidence	Higher acceptance may remain inappropriate	n
<i>Workflow-disruption disuse</i>	Poor timing and high interaction cost	Reduced usefulness and explanation fatigue	Bypass or abandonment	Readiness and design fit affect sustained use	Integrate concise explanations at decision points	Reduced detail may impair scrutiny	p
<i>Expertise-enabled critical reliance</i>	Domain knowledge and opportunity to investigate disagreement	Critical interpretation and mental-model formation	Calibrated reliance or principled rejection	Experts compare AI advice with other evidence when practices permit	Preserve authority to challenge and examine supporting evidence	Expertise may produce overconfidence	e n

Expertise-limited false reassurance	Limited AI or domain literacy	Cognitive mismatch and perceived transparency	Misuse or misplaced confidence	Accessible explanations may still fail to communicate limitations	Use progressive disclosure, uncertainty, and comprehension checks	Fluent interfaces may remain persuasive	e
Risk-sensitive cautious reliance	High-consequence and uncertain decision	Uncertainty interpretation	Deferral, verification, or conditional use	High-risk settings require stronger evidence and limitation communication	Present alternatives and escalation thresholds	Uncertainty may prompt indiscriminate rejection	
Weak-governance accountability gap	Diffuse responsibility and nominal oversight	Unclear answerability	Symbolic transparency and misuse	Visibility does not create institutional accountability	Map system ownership, review duties, and incident handling	Formal governance may remain performative	E
Contestability-enabled correction	Consequential decision with accessible review	Actionability and contestability awareness	Challenge, correction, or recourse	Explanations become consequential when linked to feasible action	Pair reasons with human review and status tracking	Recourse may remain costly or inaccessible	e w
Feedback-enabled learning	Repeated use and outcome feedback	Mental-model updating	Improved use, reporting, or redesign	Calibration can change through observed errors and outcomes	Build reporting, review, and redesign loops	Feedback may be ignored or selectively captured	
No universal pathway	Changed domain, user, system, incentive, or risk	Mechanism may fail to activate or reverse	Different or opposing outcome	Effects vary across tasks and stakeholders	Treat configurations as locally testable propositions	Overgeneralization from prominent domains	r I ir

Figure 2 presents the realist synthesis model of explainable AI use, linking context-sensitive explanation design, trust mechanisms, human responses, adoption pathways, misuse risks, decision outcomes, and implementation feedback.

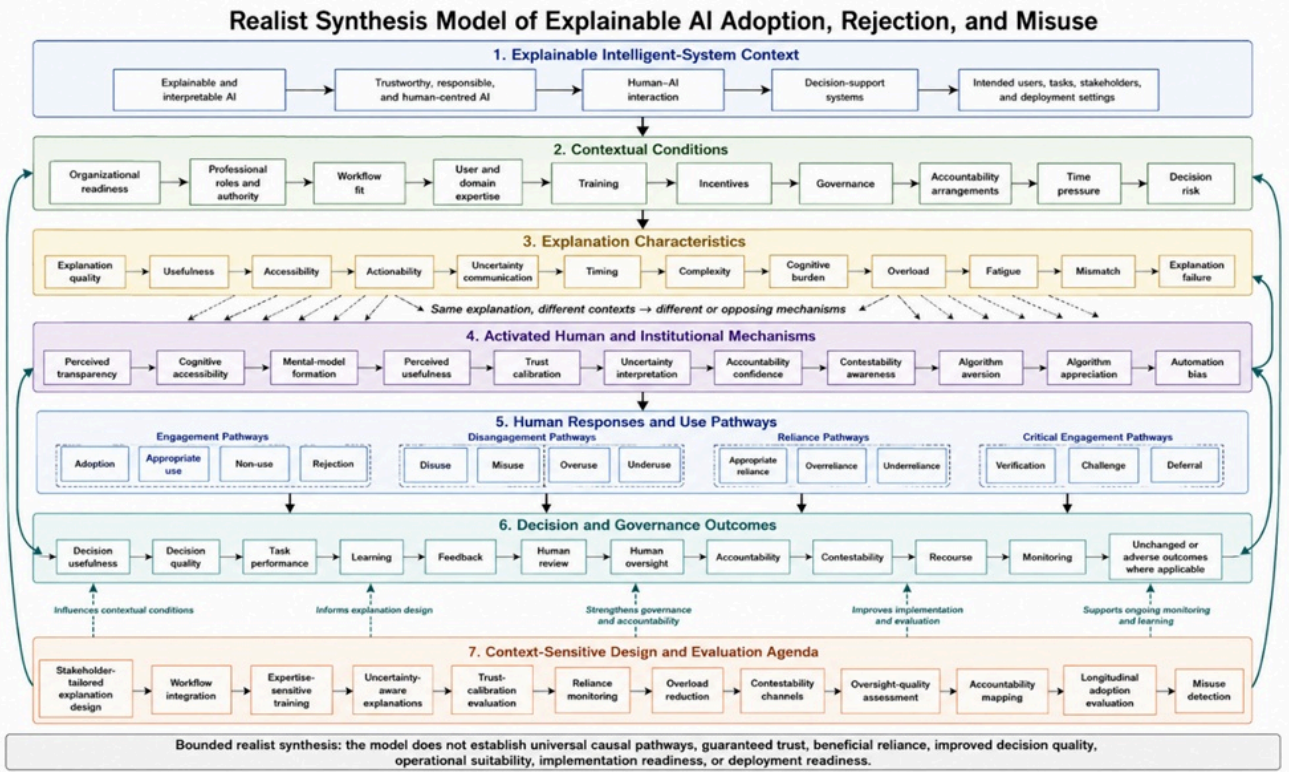


Figure 2. Realist Synthesis Model of Explainable AI Adoption, Rejection, and Misuse

Alt-text description:

A conceptual realist synthesis model showing explainable AI use across contexts, mechanisms, human responses, adoption pathways, misuse risks, decision outcomes, design agenda, and feedback loops.

Design and Implementation Agenda

Explanation design should begin with the user's role, decision purpose, expertise, and capacity to act rather than with the explanation method alone. Stakeholder-sensitive design requires separating developer debugging, professional verification, managerial accountability, auditor traceability, and affected-person recourse. Progressive disclosure may reduce overload while preserving access to technical depth where needed [7].

Explanations should also be integrated into actual workflows and accompanied by task-specific training. Evaluation should examine whether users notice limitations, detect contradictions, interpret uncertainty, and distinguish model competence from persuasive presentation. Human-centered evaluation must therefore extend beyond satisfaction and include comprehension, reliance behavior, workload, and decision consequences [11].

Trust-calibration assessment should compare confidence and advice-taking with system performance and contextual evidence. Monitoring should identify overreliance, underreliance, automation bias, overrides, out-of-scope use, and cases in which users accept AI recommendations despite contradictions. Explanation interfaces should reduce unnecessary complexity without presenting simplified narratives as proof of validity [21].

Implementation also requires contestability channels, substantive human oversight, organizational accountability mapping, incident review, and longitudinal assessment of adoption, workarounds, disuse, learning, and misuse. These

priorities are realist-review implications and research directions, not procurement, legal, compliance, certification, model-selection, platform-selection, workflow, safety, or deployment instructions [30].

Conclusion

Explainable-AI adoption is best understood as a context-dependent process rather than a direct consequence of explanation availability. Organizational readiness, professional authority, workflow fit, user expertise, technical quality, decision risk, incentives, and governance conditions shape whether explanations activate understanding, usefulness, calibrated trust, contestability, aversion, automation bias, or cognitive burden. These mechanisms can lead to adoption, careful use, non-use, rejection, disuse, overuse, underuse, misuse, learning, challenge, or accountability outcomes. The realist synthesis therefore shifts evaluation away from asking whether an explanation exists and toward examining who receives it, for what purpose, under which constraints, with what opportunity to verify or contest it, and with what decision consequences. The proposed configurations and design priorities remain bounded explanatory propositions requiring context-specific human-centered validation, organizational assessment, monitoring, and outcome evaluation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Sep 2025

Revised: 19 Oct 2025

Accepted: 17 Dec 2025

Published online: 15 March 2026

Rights and permissions

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

References

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):93:1-93:42.
<https://doi.org/10.1145/3236009>.

Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access.* 2018;6:52138-60.
<https://doi.org/10.1109/ACCESS.2018.2870052>.

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>.

Jiang J, Kahai S, Yang M. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *Int J Hum Comput Stud.* 2022;165:102839.
<https://doi.org/10.1016/j.ijhcs.2022.102839>.

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>.

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>.

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):24:1-24:45.
<https://doi.org/10.1145/3387166>.

Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion.* 2021;76:89-106.
<https://doi.org/10.1016/j.inffus.2021.05.009>.

Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput Surv.* 2023;55(13s):295:1-295:42.
<https://doi.org/10.1145/3583558>.

Vaccaro M, Almaatouq A, Malone TW. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat Hum Behav.* 2024;8(12):2293-303.
<https://doi.org/10.1038/s41562-024-02024-1>.

Jöhnk J, Weißert M, Wyrski K. Ready or not, AI comes—An interview study of organizational AI readiness factors. *Bus Inf Syst Eng.* 2021;63(1):5-20.
<https://doi.org/10.1007/s12599-020-00676-7>.

Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ Sci*. 2022;33(1):126-48.
<https://doi.org/10.1287/orsc.2021.1549>.

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>.

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).

Kim M, Kim S, Kim J, Song TJ, Kim Y. Do stakeholder needs differ? Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. *Int J Hum Comput Stud*. 2024;181:103160.
<https://doi.org/10.1016/j.ijhcs.2023.103160>.

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>.

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>.

Papenmeier A, Kern D, Englebienne G, Seifert C. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Trans Comput Hum Interact*. 2022;29(4):35:1-35:33.
<https://doi.org/10.1145/3495013>.

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

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

Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: People will use imperfect algorithms if they can—even slightly—modify them. *Manage Sci*. 2018;64(3):1155-70.
<https://doi.org/10.1287/mnsc.2016.2643>.

Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. *Organ Behav Hum Decis Process*. 2019;151:90-103.
<https://doi.org/10.1016/j.obhdp.2018.12.005>.

Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. *J Mark Res*. 2019;56(5):809-25.
<https://doi.org/10.1177/0022243719851788>.

Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. *J Behav Decis Mak*. 2020;33(2):220-39.
<https://doi.org/10.1002/bdm.2155>.

Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Comput Hum Behav*. 2024;160:108352.
<https://doi.org/10.1016/j.chb.2024.108352>.

Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput Surv.* 2022;55(5):95:1-95:29.

<https://doi.org/10.1145/3527848>.

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>.

Shin D, Park YJ. Role of fairness, accountability, and transparency in algorithmic affordance. *Comput Human Behav.* 2019;98:277-84.

<https://doi.org/10.1016/j.chb.2019.04.019>.

Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res.* 2019;46(4):629-50.

van der Waa J, Nieuwburg E, Cremers A, Neerincx 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>.