

ORIGINAL RESEARCH

Open access

Explanation Is Not Enough—A Conceptual Framework for Designing AI Decisions That People Can Challenge, Override, Appeal, and Correct

Elena Petrova¹, Marcus Weber^{2*}, Priya Nair³

Abstract

Artificial intelligence increasingly contributes to decisions that shape access, opportunity, services, evaluation, and institutional treatment. Although explainability is widely treated as a foundation of trustworthy AI, an explanation does not by itself enable an affected person to question a decision, introduce corrective evidence, request reconsideration, or obtain an appropriate response. This article develops a conceptual Human Contestability Framework for connecting explanation to actionable decision challenge. The framework treats contestability as a socio-technical requirement involving affected-person access, reason-giving, explanation quality, procedural accessibility, evidence submission, meaningful human review, override capacity, escalation, appeal, redress, correction, organizational accountability, documentation, auditability, monitoring, and feedback updating. It distinguishes interpretability and transparency from contestability, and separates human review from meaningful reconsideration, appeal from recourse, and correction from broader redress. The framework proposes that contestable decision design should begin with decision-context and stakeholder mapping, define explanation and evidence requirements, establish challenge and review pathways, assign organizational responsibilities, preserve traceable records, and use contestation outcomes to identify recurring system or governance failures. Its contribution is not a validated procedure or universal institutional model, but an integrated structure for analyzing whether people can move from receiving an AI-supported decision to questioning its basis and seeking a reasoned institutional response. Contestability may strengthen trustworthy-AI planning, but it does not independently establish fairness, legality, ethical sufficiency, accountability, operational safety, or deployment readiness. Domain-specific validation, accessible procedures, competent human review, organizational governance, and continuing evaluation remain necessary.

Keywords Explainable AI, Trustworthy AI, Human override, Redress, Contestable AI, Appeals

*Correspondence:

Marcus Weber
marcus.weber@lmu.de

¹ Department of Trustworthy AI and Governance, Faculty of Computer Science, Sofia University, Bulgaria

² Department of AI Ethics and Reliability Engineering, Faculty of Informatics, LMU Munich, Munich, Germany

³ Department of Responsible AI and Human-Centered Computing, Indian Institute of Technology Bombay, Mumbai, India

Introduction

Explainable artificial intelligence encompasses diverse technical methods, explanation purposes, audiences, and interpretive assumptions. Human explanation research further shows that useful explanations are often selective, contrastive, socially situated, and responsive to the questions people actually ask, while technical surveys demonstrate that explanation methods differ substantially in scope, fidelity, and output [1-3]. These differences matter when AI

contributes to consequential decisions because a technically available explanation may still be inaccessible, incomplete, unresponsive, or disconnected from any process through which its reasoning can be challenged.

Trustworthy-AI discourse commonly connects transparency with fairness, responsibility, accountability, and human control, yet high-level principles do not automatically become usable procedures for people affected by algorithmic decisions [4]. A person may receive a feature-attribution display, confidence score, or generic description of model logic without knowing which evidence was used, which uncertainties mattered, who owns the decision, how to submit corrective information, or whether a reviewer can reconsider the outcome. Explanation therefore provides only one component of a wider contestability architecture.

Human contestability concerns the practical capacity to question a decision, dispute its factual or normative basis, provide relevant evidence, request meaningful reconsideration, escalate unresolved concerns, and receive a reasoned response. Depending on context, it may also include override, appeal, correction, recourse, or redress. These functions are related but not interchangeable. Questioning may clarify a decision without changing it; reconsideration reopens its substance; override interrupts or replaces an output; appeal introduces a further review level; correction repairs an identified error; and redress addresses relevant consequences where institutionally supported.

This article proposes a Human Contestability Framework that links explanation, decision challenge, human review, appeal, correction, accountability, and organizational learning. The framework is intended to clarify conceptual relationships, required inputs, responsible functions, procedural dependencies, evidence infrastructures, and residual risks. It does not prescribe a universal appeals mechanism, establish legal or procedural sufficiency, or imply that every AI-supported decision should use the same form of review. Its contribution is a structured basis for designing and evaluating challengeable, overrideable, appealable, and correctable decision processes without treating any single safeguard as sufficient.

Contestability as a Requirement of Trustworthy AI

Contestability extends trustworthy-AI design beyond disclosure by connecting system explanations to affected-person participation, institutional answerability, and pathways for corrective action. Contestable-AI scholarship frames contestability as a lifecycle property rather than a post-decision interface; normative analysis emphasizes that decision subjects may be owed opportunities for participation and challenge beyond transparency; and accountability scholarship links reason-giving to identifiable relationships of answerability and consequence [5-7]. Contestability is therefore best understood as a socio-technical capacity distributed across interfaces, reviewers, organizational roles, records, and governance arrangements.

At minimum, a contestable process requires notice of the decision, access to decision-relevant information, an understandable challenge route, an opportunity to submit evidence, a responsible reviewing function, and a reasoned response. Human involvement alone cannot satisfy these conditions. Human-oversight policies may fail when reviewers lack time, competence, authority, independence, institutional support, or realistic opportunities to depart from automated recommendations [8]. Meaningful contestability must consequently evaluate what humans can actually examine and change rather than merely confirm that a human is formally present.

Procedural accessibility is central because a challenge mechanism that cannot be discovered, understood, afforded, or completed is only nominally available. Affected-person access requires clear communication about who may raise concerns, what elements can be questioned, which evidence may be submitted, how the case will progress, and when a response should be expected. Decision-subject access additionally requires information specific to the individual decision rather than only general model documentation. Escalation should be possible when an initial reviewer lacks authority, competence, independence, or capacity to resolve the disputed issue.

Contestability should nevertheless not be treated as a guarantee of trustworthy outcomes. A procedurally accessible challenge can still operate within a defective decision system, reproduce institutional power imbalances, or offer remedies too limited to address material harm. Explanation may support contestability but cannot establish it alone; contestability may support accountability but cannot certify it; and a completed review pathway does not prove fairness, compliance, safety, or deployment readiness. The framework therefore incorporates explicit no-sufficiency and no-guarantee boundaries alongside its positive procedural requirements.

Explanations for Decision Challenge

Explanations intended for contestability must be evaluated according to the task they are meant to support. Explanation quality may involve fidelity, correctness, relevance, completeness, stability, and comprehensibility, while usefulness concerns whether the explanation helps a person identify a disputed input, assumption, rule, uncertainty, or decision boundary. Existing XAI evaluation approaches remain heterogeneous, making it unsafe to infer that an explanation is useful merely because it is available or positively received [9]. Contestability consequently requires purpose-specific evaluation tied to questioning, evidence submission, and reconsideration.

Static disclosure may also be insufficient when the affected person cannot ask follow-up questions. Conversational and layered explanation models suggest that people may need progressively deeper information, including why an outcome occurred, why an alternative did not occur, what evidence mattered, and how contextual assumptions affected the decision [10]. Such interaction should not be confused with persuasion. Its contestability value lies in helping a person identify what can be questioned, which information may be missing, and whether a formal review route is appropriate.

Explanation accessibility depends on language, format, cognitive load, technical literacy, domain knowledge, disability access, and the person's uncertainty about the system or decision. Empirical evidence indicates that explanation demand varies with users' epistemic uncertainty and that additional explanation is not uniformly beneficial [11]. A contestable system should therefore permit layered detail and supportive interpretation without assuming that personalization automatically produces understanding. User confidence, readability, or satisfaction cannot substitute for direct assessment of comprehension and challenge performance.

Decision rationales should identify the material grounds for the outcome, while evidence-basis explanations should describe the relevant data, records, assumptions, provenance, and limitations. Feature contributions may help identify influential inputs where technically appropriate, but they should not be presented as complete causal reasons. Counterfactual and contrastive explanations can clarify decision boundaries, yet algorithmic-recourse research shows that apparent actionability may be infeasible, causally invalid, discriminatory, unstable, or unfairly burden-shifting [12]. Institutional correction and redress must therefore remain distinct from recommendations that a person alter their circumstances.

Actionable explanation should primarily clarify what may be questioned, corrected, supplemented, or escalated within the decision process. Domain experts may be needed to interpret specialized evidence, while affected persons require a form that remains procedurally and cognitively accessible. Explanation design should disclose uncertainty and limitations, connect technical information to decision-specific reasons, and identify available next steps without implying guaranteed reversal or remedy. **Table 1** summarises explanation requirements for contestable AI decisions, linking reason-giving, evidence basis, accessibility, uncertainty, actionability, and challenge pathways.

Table 1. Explanation Requirements for Human Contestability in AI Decisions: Reason-Giving, Explanation Quality, Explanation Usefulness, Accessibility, Completeness, Uncertainty, Evidence Basis, Counterfactual and Contrastive Support, Actionability, Cognitive Accessibility, Decision Challenge, and Explanation-Sufficiency Boundaries

Explanation or contestability requirement	Core challenge question	Required explanation feature	Contestability function	Potential output	Failure or misinterpretation risk	Human interpretation or governance need
Explanation quality	Is the explanation accurate, faithful, relevant, and sufficiently stable?	Method-appropriate fidelity, correctness, and relevance checks	Determines whether the explanation can support scrutiny	Evaluated explanation artifact	Persuasive but unfaithful narrative	Technical and human-centered validation
Explanation usefulness	Does the information help identify a challengeable issue?	Task-specific and decision-relevant content	Connects understanding to questioning	Identified challenge ground	Equating satisfaction with usefulness	Challenge-tailored evaluation
Accessibility and cognitive accessibility	Can the affected person perceive and understand the explanation?	Plain language, accessible formats, layering, and support	Enables practical use	Accessible explanation package	Technical disclosure that transfers burden to the person	Inclusive comprehensibility testing
Completeness and uncertainty	Are material reasons, limitations, omissions, and uncertainties disclosed?	Decision-relevant coverage and calibrated uncertainty	Reveals disputed assumptions and evidence gaps	Completeness and uncertainty statement	Information overload, omission, or false precision	Layered communication and risk interpretation
Reason-giving and decision rationale	Why was the decision reached?	Decision-specific reasons, criteria, and contextual judgment	Supports direct questioning	Reason statement	Post-hoc rationalization	Fidelity and domain review
Evidence basis	What data and records materially supported the decision?	Sources, provenance, material evidence, and limitations	Enables factual dispute and correction	Evidence summary	Unverifiable evidence or privacy exposure	Data-governance and access controls
Feature contribution where supported	Which inputs influenced the model output?	Valid local or global attribution with limitations	Identifies potentially disputed inputs	Contribution display	Treating association as causation or full rationale	Statistical and domain interpretation
Counterfactual explanation	What change might have	Feasible, stable, and causally	Clarifies decision boundaries	Counterfactual scenario	Impossible or burden-shifting recommendation	Feasibility and causal assessment

where supported	altered the output?	cautious alternatives				
Contrastive explanation where supported	Why this outcome rather than a relevant alternative?	Explicit comparison outcome and distinguishing reasons	Supports targeted questioning	Contrastive reason	Manipulative or irrelevant comparison	User-sensitive selection of the comparison
Actionable explanation	What can be questioned, corrected, or supplemented?	Clear distinction between procedural actions and behavioural recommendations	Links reasons to challenge and correction pathways	Challenge options	Confusing recourse with institutional remedy	Procedural guidance and governance review
User and domain expertise	What level of detail is required for meaningful interpretation?	Adjustable depth and access to qualified expertise	Supports both personal and specialist scrutiny	Expertise-appropriate explanation	Incorrect expertise assumptions or inaccessible technical review	User control and competent reviewer access
Decision challenge pathway	How can the explanation be used to initiate review?	Visible route for questioning, evidence submission, status, and escalation	Converts explanation into contestability	Registered challenge	Disclosure without usable procedure	End-to-end pathway testing
No explanation-sufficiency boundary	Is explanation connected to review and correction?	Challenge, reconsideration, response, escalation, and correction options	Prevents explanation-only governance	Explicit pathway linkage	Claiming clarity itself establishes contestability	Framework integrity assessment
No human-understanding guarantee boundary	Has understanding been assessed rather than presumed?	Direct comprehension and task-performance evidence	Limits unsupported claims	Comprehension evidence with limitations	Treating readability or confidence as understanding	Human-centered evaluation

Human Override and Decision Reconsideration

Human review should mean substantive examination of the disputed decision rather than ceremonial confirmation of an AI output. Effective oversight requires clarity about what is reviewed, when intervention can occur, who holds authority, and how responsibility is allocated [13]. The reviewer should have access to the original decision record, relevant explanations, submitted evidence, system limitations, and applicable decision criteria. Without these inputs, the reviewer may be unable to distinguish model error, data error, contextual omission, procedural failure, or a legitimate disagreement about judgment.

Meaningful human review additionally requires practical capacity to question and alter the outcome. Human intervention should not function as a formal step through which institutions disclaim responsibility while preserving automated

conclusions [14]. Reviewers need sufficient competence, time, independence, organizational support, and authority to confirm, modify, reverse where supported, suspend, or escalate a decision. Their reasoning and use of evidence should be documented so that the review process itself remains open to later examination.

Human control should also be interpreted beyond the immediate operator. A society-in-the-loop perspective recognizes that consequential systems embody institutional objectives and value choices that may require collective or governance-level reconsideration, not only case-by-case intervention [15]. Individual review can correct a particular decision, while governance review may identify recurring explanation failures, inappropriate decision boundaries, systematic data problems, unsuitable escalation criteria, or institutional assumptions that cannot be resolved through frontline override alone.

Neither human-in-the-loop nor human-on-the-loop arrangements inherently provide superior or fairer decisions. Experimental evidence indicates that human decision makers are not automatically perceived as fairer than machine decision makers, reinforcing the need to evaluate the actual quality of review rather than relying on decision-maker identity [16]. Human reviewers may introduce automation bias, inconsistency, fatigue, unsupported discretion, or institutional defensiveness. Reviewer presence must therefore be assessed alongside independence, competence, authority, reasoning quality, workload, and the consequences of intervention.

Override capacity should be distinguished from meaningful control. A technically available override mechanism may be unusable when intervention is too late, risks are unclear, authority is ambiguous, or organizational incentives discourage departure from automated recommendations. Review documentation and review logs should record material evidence, questions, actions, reasoning, timing, and outcome without being treated as proof that review was adequate. Human review does not guarantee fairness, and override does not certify accountability, safety, or control; both require contextual evaluation, monitoring, and governance [8].

Appeals, Redress, and Procedural Fairness

Appeal, redress, recourse, and correction address different dimensions of contestability. An appeal requests further review of an earlier determination, while recourse concerns feasible actions that may improve a person's position. Correction repairs erroneous data, reasoning, procedure, or outcomes, whereas redress responds to relevant consequences where an institution has authority and capacity to do so. Fairness-perception research indicates that people evaluate algorithmic decisions through both outcome-related and procedural considerations, including transparency, voice, consistency, and the identity of the decision maker [17]. These perceptions are important for design evaluation but cannot independently establish substantive fairness.

A usable appeals pathway requires decision-specific access to information. Affected persons need sufficiently intelligible information about the logic, significance, evidence, and anticipated consequences of an automated decision before they can identify meaningful grounds for challenge [18]. Process notice should clarify available stages, relevant timelines, responsible functions, evidentiary opportunities, possible outcomes, and escalation routes. Access to general system documentation alone may be inadequate when the disputed issue concerns individualized evidence, contextual circumstances, or a particular combination of human and automated judgment.

A contestable pathway should connect explanation and questioning to evidence submission, human reconsideration, escalation, appeal, reasoned response, correction, and organizational learning. Procedural fairness may be supported when affected stakeholders can express concerns and when institutions use structured processes to identify and reason through competing values; nevertheless, participation does not prove fair outcomes, and no single deliberative structure is universally appropriate [19, 20]. **Figure 1** illustrates the human contestability pathway, showing how an AI decision can move from explanation and challenge to human reconsideration, appeal, redress, correction, accountability, and feedback updating.

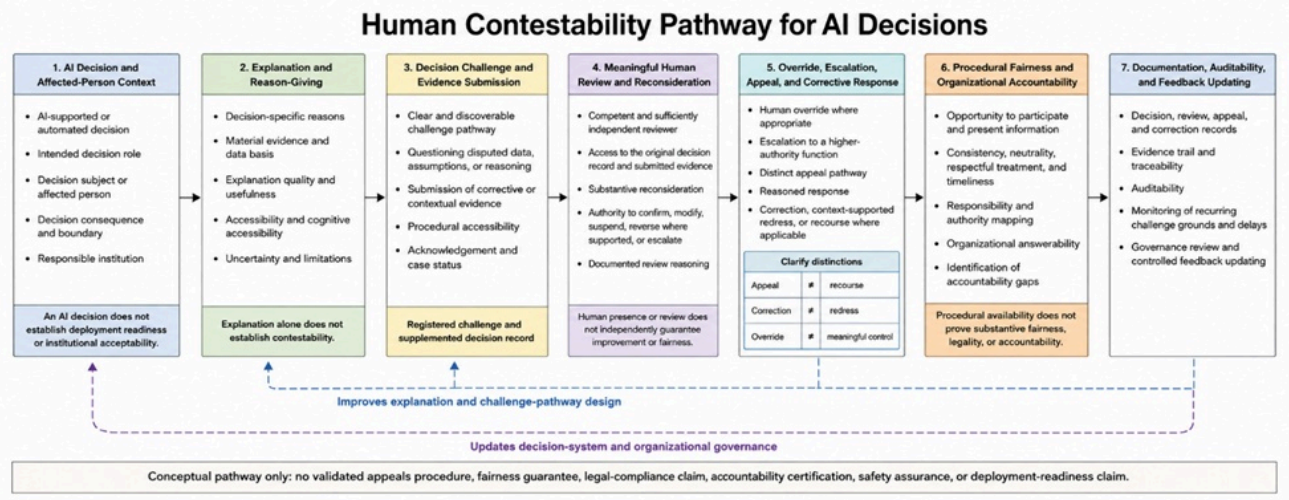


Figure 1. Human Contestability Pathway for AI Decisions

Alt-text description:

A conceptual pathway showing AI decision explanation, affected-person challenge, evidence submission, human review, reconsideration, override, appeal, redress, correction, accountability, documentation, and feedback updating.

Procedural fairness concerns the qualities of the process through which a decision is reached and reconsidered. Relevant considerations may include procedural accessibility, opportunity to present information, neutrality, consistency, respectful treatment, timeliness, and reason-giving. Procedural justice additionally directs attention to the legitimacy of institutional authority and the treatment of competing interests. Due-process terminology should be used only where conceptually and contextually supported; the framework does not infer legal entitlements or prescribe jurisdiction-specific hearing requirements.

A reasoned response should address the material questions raised, identify the evidence considered, explain the resulting determination, communicate uncertainties or limitations, and state any remaining options. Successful correction should extend beyond changing a displayed decision when the original error has propagated into data records, downstream systems, or organizational practices. Appeal and correction outcomes should therefore feed controlled monitoring and root-cause analysis. Appeals do not guarantee fairness, redress cannot eliminate every consequence, and procedural availability alone does not establish legal or ethical sufficiency.

Organizational Responsibilities for Contestability

Contestability depends on organizational governance because explanations, reviews, appeals, corrections, and monitoring require authority, resources, and coordinated ownership. Responsible-AI governance research indicates that organizations need linked structures, processes, roles, controls, and lifecycle mechanisms rather than isolated ethical principles [21]. Institutional accountability should identify which organization remains answerable for the decision, while organizational accountability should specify which units own explanation provision, review, correction, escalation, record preservation, and systemic learning.

AI governance is distributed across developers, deployers, data stewards, operators, decision makers, reviewers, suppliers, assurance functions, and affected stakeholders [22]. Responsibility mapping should distinguish these roles and reveal points at which authority or answerability may disappear. Accountability-gap analysis is particularly important where one actor develops the model, another supplies data, a third integrates the system, and a separate institution makes or communicates the decision. Shared responsibility should not be interpreted as the absence of responsibility.

Data governance supports contestability by preserving information about provenance, quality, transformations, access, ownership, and permitted use [23]. A reviewer cannot reliably examine a disputed decision when relevant data cannot be traced or when derived variables, supplier inputs, and processing histories are unavailable. Decision documentation should record material inputs, outputs, human contributions, uncertainty, and final status. Model and data documentation should provide context for technical scrutiny without being mistaken for individualized decision explanations.

High-level transparency principles must also be translated into explicit requirements that identify audiences, information needs, responsible actors, system components, and evaluation criteria [24]. Reviewer mandates should specify scope, authority, competence expectations, conflict rules, and access to specialized support. Decision, review, appeal, and correction logs should preserve relevant actions, evidence, reasoning, timing, versions, and implementation status. Auditability requires that authorized evaluators can inspect this evidence, whereas traceability concerns the ability to connect decisions, models, data, actors, and changes across the lifecycle.

Monitoring should examine recurring challenge grounds, procedural delays, inaccessible pathways, reversal or correction patterns, unresolved harms, and differential effects across relevant groups. Governance review can use these signals to determine whether problems arise from explanation design, data quality, model limitations, reviewer practice, institutional rules, or the decision context itself. Documentation and logs may support accountability, but they cannot certify it without answerability, authority, correction capacity, and consequential follow-through [7]. **Table 2** maps human override, decision reconsideration, appeals, redress, procedural fairness, and organizational accountability requirements for contestable AI decision systems.

Table 2. Human Override, Decision Reconsideration, Appeals, Redress, Procedural Fairness, and Organizational Accountability Requirements for Contestable AI: Human Review, Override Capacity, Reviewer Independence, Appeal Pathways, Correction, Reasoned Response, Auditability, Traceability, Documentation, Review Logs, Appeal Logs, and Accountability Boundaries

Contestability or accountability requirement	Core procedural or governance question	Responsible actor or process	Required evidence or practice	Contestability or accountability function	Potential output	Failure or responsibility gap
Meaningful human review	Can a human substantively examine and alter the decision?	Competent, empowered reviewer	Decision record, explanation, evidence, time, authority, and review criteria	Reconsiders the decision	Confirmed, modified, reversed where supported, or escalated decision	Rubber-stamping or automation bias
Human oversight	Is system operation monitored across the decision lifecycle?	Operator, supervisor, or governance function	Monitoring indicators, intervention criteria, logs, and escalation rules	Detects individual and systemic failures	Intervention or governance referral	Oversight without authority or resources
Human override and override capacity	Can an authorized person stop or replace an AI	Designated decision authority	Usable control, defined grounds, competence, authority, and fallback process	Enables immediate intervention	Overridden or suspended output	Technical control without practical ability to use it

Reasoned response	Does the institution explain how the challenge was considered?	Reviewer or responsible decision function	Issues raised, evidence considered, reasoning, outcome, uncertainty, and next steps	Demonstrates answerability	Accessible response	Generic response that ignores material arguments
Procedural fairness	Is the process accessible, participatory, consistent, impartial, and reasoned?	Review and governance functions	Voice, neutrality, respectful treatment, consistency, and timeliness	Evaluates procedural quality	Procedural assessment	Treating perceived fairness as substantive fairness
Responsibility mapping	Who must explain, review, correct, escalate, and learn?	Organizational governance function	Role–task–authority map	Prevents responsibility diffusion	Responsibility allocation	Named roles without authority or resources
Decision, model, and data documentation	Can the decision and relevant technical context be reconstructed?	Decision owner, developer, and data steward	Inputs, outputs, versions, provenance, limitations, human actions, and changes	Supports scrutiny	Linked documentation package	Generic records disconnected from the decision
Decision, review, appeal, and correction logs	Can procedural events and actions be traced?	System, records, review, and appeals functions	Actors, evidence, actions, reasoning, timing, outcome, and implementation	Supports monitoring and audit	Linked case history	Incomplete, formulaic, or inaccessible logs
Auditability and traceability	Can authorized evaluators inspect duties and reconstruct dependencies?	Assurance and governance functions	Evidence integrity, access controls, identifiers, lineage, and retention	Enables structured examination	Audit findings or trace reconstruction	Treating logs as proof of accountability
Monitoring and governance review	Are recurring failures identified and addressed institutionally?	Operational monitoring and governance functions	Trends, delays, outcomes, subgroup patterns, root causes, and action ownership	Supports systemic correction	Governance determination and controlled update	Individualizing structural problems
No accountability guarantee boundary	Do records and roles produce real answerability	Institutional governance	Evidence of authority, consequence, response, and implementation	Limits accountability overclaiming	Residual accountability-gap statement	Paper accountability without practical consequence

	and corrective capacity?				
--	--------------------------	--	--	--	--

Proposed Human Contestability Framework

The proposed framework begins with use-case and decision-context definition because contestability requirements depend on what decision is being made, how AI contributes to it, who exercises authority, and what consequences may follow. Ethics-to-practice scholarship shows that principles require lifecycle tools and methods, although the availability of such instruments does not itself establish responsible practice [25]. Decision-context definition should therefore identify the system’s role, human contribution, affected processes, foreseeable consequences, decision boundaries, and institutional ownership before a challenge mechanism is specified.

The second stage maps affected persons and stakeholders and assesses the potential severity, reversibility, scale, recurrence, uncertainty, and distribution of decision impacts. Responsible design should attend to socially significant objectives, relevant stakeholders, feasibility, and foreseeable harm rather than treating technical optimization as the sole design problem [26]. Impact classification is a contextual planning device, not a certification score. It helps determine the required depth of explanation, urgency of review, independence of reconsideration, and strength of documentation and monitoring.

The third and fourth stages define explanation requirements and connect them to challenge, evidence submission, review, and escalation. Contemporary XAI research emphasizes the need for interdisciplinary, interactive, context-sensitive, and rigorously evaluated explanations [27]. The framework therefore requires explanation purpose, audience, evidence basis, uncertainty, accessibility, and limitations to be specified before selecting an explanation method. Challenge interfaces should allow disputed facts, assumptions, data, or reasoning to be identified and should preserve submitted evidence for meaningful reconsideration.

The fifth stage integrates human review, override, appeal, redress, correction, and procedural-fairness evaluation. End-user XAI research indicates that explanation needs concern format, completeness, accuracy, and currency as well as technical content [28]. The review pathway should therefore distinguish clarification from reconsideration, and reconsideration from appeal. Override should remain conditional on authority and context; correction should include relevant downstream records; and redress should not be inferred merely because the original decision has been changed.

The final stages assign organizational responsibilities, design documentation and auditability, link decision, review, appeal, and correction logs, and establish monitoring and governance feedback. Human-centered AI treats automation and human control as dimensions that can be jointly designed rather than opposing endpoints [29]. The framework accordingly combines technical, procedural, human, and institutional controls while preserving explicit boundaries against explanation sufficiency, human-review sufficiency, override sufficiency, appeal-fairness guarantees, redress guarantees, accountability certification, compliance claims, and deployment-readiness claims. **Table 3** presents the proposed human contestability framework linking decision context, affected-person access, explanations, challenge pathways, human review, override, appeal, redress, correction, accountability, auditability, monitoring, and feedback updating.

Table 3. Proposed Human Contestability Framework for AI Decisions: Decision Context, Affected-Person Mapping, Explanation Requirements, Reason-Giving, Evidence Basis, Challenge Pathways, Evidence Submission, Human Review, Reconsideration, Override, Appeal, Redress, Correction, Procedural Fairness, Organizational Accountability, Auditability, Monitoring, Feedback Updating, and Deployment Boundaries

Framework stage	Core purpose	Required input	Contestability function	Potential output	Validation or governance	Failure risk
-----------------	--------------	----------------	-------------------------	------------------	--------------------------	--------------

					need	
Use-case and decision-context definition	Define what decision is made, by whom, and with what consequences	Purpose, workflow, system role, institutional authority, and downstream effects	Establishes the object of contestation	Decision-context statement	Domain and institutional review	Vague scope or concealed secondary use
Affected-person and stakeholder mapping	Identify people with contestability needs and actors with duties	Direct and indirect affected persons, users, operators, developers, reviewers, and governance actors	Defines access, participation, and responsibility	Stakeholder map	Participatory and impact-sensitive assessment	Exclusion of low power or indirectly affected groups
Decision-impact and risk classification	Match procedural depth to potential consequences	Severity, reversibility, scale, vulnerability, recurrence, and uncertainty	Informs explanation, review, and escalation intensity	Contextual impact classification	Periodic domain reassessment	Underclassification or false precision
Explanation-requirement definition	Specify the information required for questioning	Stakeholder needs, decision logic, risk, expertise, and available methods	Defines explanation purpose and audience	Explanation specification	Human-centered evaluation	Generic transparency requirements
Reason-giving and evidence-basis design	Connect the outcome to material reasons and evidence	Rationale, data provenance, model contribution, human judgment, and uncertainty	Enables factual and reasoning challenge	Decision-specific reason package	Fidelity, privacy, completeness, and domain review	Plausible but unfaithful reasoning
Challenge-pathway design	Define how a person initiates and follows a challenge	Entry point, grounds, stages, owners, timing, status, and escalation	Converts disagreement into a managed process	Registered challenge	Accessibility and end-to-end testing	Hidden or fragmented pathway
Evidence-submission mechanism	Permit relevant corrective and contextual information	Accepted evidence, secure channel, provenance, relevance criteria, and acknowledgement	Supplements the decision record	Evidence-submission record	Privacy, security, and evidence-handling controls	Evidence gathered but not considered
Human review and	Provide substantive	Original record, explanation,	Reopens the decision	Reasoned review	Competence, independence,	Rubber-stamping or automation bias

reconsideration	reassessment	challenge, evidence, and criteria		outcome	workload, and authority assessment	
Override and escalation design	Enable urgent intervention or higher-authority review	Override grounds, authority, control point, fallback, and escalation criteria	Prevents automatic continuation or dead-end review	Override, suspension, or escalation record	Operational and governance evaluation	Unusable or arbitrary intervention
Appeal-pathway design	Provide a distinct further-review route	Eligibility, timelines, record transfer, authority, independence, and communication	Supports further reconsideration	Appeal determination	Accessibility, consistency, and implementation monitoring	Same reasoning repeated under new label
Redress and correction design	Address verified error and relevant consequences	Error, harm, authority, downstream effects, and contextual response options	Repairs records or supports an appropriate response	Correction or context-supported redress	Domain-specific authorization and verification	Symbolic response or incomplete propagation
Procedural-fairness review	Evaluate process quality	Access, participation, neutrality, reasons, timing, outcomes, and subgroup evidence	Assesses whether the pathway operates fairly in procedure	Procedural findings	Independent human-centered evaluation	Confusing perceived and substantive fairness
Organizational-responsibility mapping	Assign explanation, review, correction, escalation, and learning duties	Workflow, roles, authority, suppliers, and governance structure	Prevents responsibility diffusion	Accountability map	Leadership authorization and gap analysis	Roles without authority or resources
Documentation and auditability design	Preserve evidence for reconstruction and scrutiny	Decision, model, data, process, version, and role requirements	Enables examination	Documentation architecture	Integrity, access, privacy, security, and retention controls	Excessive, missing, or inaccessible records
Review, appeal, and correction logging	Record actions and outcomes	Actors, evidence, reasons, actions, timing, implementation, and status	Supports traceability and systemic analysis	Linked case history	Proportional access and integrity controls	Unstructured or privacy-invasive logging

Monitoring and feedback updating	Identify recurring failures and support controlled improvement	Challenge grounds, delays, outcomes, disparities, corrections, and root causes	Converts contestation into lifecycle learning	Controlled organizational or system update	Change control and post-update evaluation	Learning from biased or manipulated feedback
Governance review	Address systemic or high-impact issues	Aggregated evidence, stakeholder input, risk analysis, and authority	Moves from individual cases to institutional reconsideration	Governance decision and assigned action	Mandate, competence, transparency, and follow-through	Symbolic governance
Explicit framework boundaries	Prevent conceptual overclaiming	Evidence on explanation, review, appeal, correction, accountability, compliance, and readiness	Preserves cautious interpretation	Boundary and residual-risk statement	Framework-integrity review	Treating framework completion as certification

Policy, Design, and Research Implications

For AI developers and deployers, the framework suggests that technical explanation features should be designed together with procedural affordances. A feature-attribution display or counterfactual output becomes contestability-relevant only when users can identify disputed elements, communicate contextual information, and reach a responsible reviewing function. Challenge interfaces should therefore be evaluated for discoverability, accessibility, cognitive burden, evidence submission, status transparency, and escalation—not merely for aesthetic usability or explanation satisfaction [5].

For decision-support operators and human reviewers, future work should examine meaningful-review quality rather than human presence alone. Relevant dimensions include reviewer competence, independence, workload, authority, access to domain expertise, use of submitted evidence, reasoning quality, and ability to intervene before harm becomes difficult to reverse. Organizations should also study when automation bias, inconsistent discretion, or institutional incentives weaken reconsideration. These are empirical evaluation questions rather than assumptions resolved by assigning a human reviewer.

For governance bodies, assurance functions, and complaint handlers where supported, contestation records may provide evidence of recurring explanation failures, inaccessible processes, data-quality problems, unresolved responsibility gaps, and unequal procedural burdens. Responsible-AI governance research supports examining such evidence across organizational structures and lifecycle processes [21]. Auditability and traceability can support this examination, but audit findings should not be interpreted as automatic proof of fairness, accountability, compliance, or safety.

Research should investigate how affected persons and decision subjects understand explanations, discover challenge pathways, submit evidence, experience review, and evaluate reasoned responses across domains and populations. It should also assess appeal accessibility, correction propagation, residual harm, stakeholder participation, and lifecycle learning [27]. These implications offer research and design directions rather than legal advice, procurement guidance, audit instructions, model-selection recommendations, prescribed grievance structures, compliance assurance, safety certification, or deployment instruction.

Conclusion

Trustworthy AI decision design should connect explanation to action. The proposed Human Contestability Framework organizes the conditions through which affected people may question AI-supported decisions, submit evidence, obtain meaningful reconsideration, trigger override or escalation where appropriate, pursue appeal, seek correction or context-supported redress, and receive a reasoned response. It also connects individual contestation to responsibility mapping, documentation, auditability, monitoring, governance review, and feedback updating. The framework clarifies that explanation, human review, override, appeal, redress, and accountability mechanisms are mutually dependent but individually insufficient. It remains a conceptual structure requiring domain-specific evaluation, accessible institutional procedures, human-centered assessment, organizational capacity, and continuing governance; it does not establish fairness, legal compliance, ethical sufficiency, operational safety, certification, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Nov 2025

Revised: 15 Dec 2025

Accepted: 04 Feb 2026

Published online: 15 March 2026

Rights and permissions

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

References

- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.
- 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>.

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>.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell.* 2019;1(9):389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.

Alfrink K, Keller I, Kortuem G, Doorn N. Contestable AI by design: towards a framework. *Minds Mach.* 2023;33(4):613-39.
<https://doi.org/10.1007/s11023-022-09611-z>.

Grant DG, Behrends J, Basl J. What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philos Stud.* 2025;182(1):55-85.
<https://doi.org/10.1007/s11098-023-02013-6>.

Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: what it is and how it works. *AI Soc.* 2024;39(4):1871-82.
<https://doi.org/10.1007/s00146-023-01635-y>.

Green B. The flaws of policies requiring human oversight of government algorithms. *Comput Law Secur Rev.* 2022;45:105681.
<https://doi.org/10.1016/j.clsr.2022.105681>.

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>.

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>.

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>.

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>.

Enqvist L. Human oversight in the EU Artificial Intelligence Act: what, when and by whom? *Law Innov Technol.* 2023;15(2):508-35.
<https://doi.org/10.1080/17579961.2023.2245683>.

Lazcoz G, de Hert P. Humans in the GDPR and AIA governance of automated and algorithmic systems: essential pre-requisites against abdicating responsibilities. *Comput Law Secur Rev.* 2023;50:105833.
<https://doi.org/10.1016/j.clsr.2023.105833>.

Rahwan I. Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf Technol.* 2018;20(1):5-14.
<https://doi.org/10.1007/s10676-017-9430-8>.

Kern C, Gerdon F, Bach RL, Keusch F, Kreuter F. Humans versus machines: who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns.* 2022;3(10):100591.
<https://doi.org/10.1016/j.patter.2022.100591>.

Starke C, Baleis J, Keller B, Marcinkowski F. Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. *Big Data Soc.* 2022;9(2):20539517221115189.
<https://doi.org/10.1177/20539517221115189>.

Custers B, Heijne AS. The right of access in automated decision-making: the scope of article 15(1)(h) GDPR in theory and practice. *Comput Law Secur Rev*. 2022;46:105727.
<https://doi.org/10.1016/j.clsr.2022.105727>.

Decker MC, Wegner L, Leicht-Scholten C. Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics Inf Technol*. 2025;27(1):1.
<https://doi.org/10.1007/s10676-024-09811-4>.

de Fine Licht K. Resolving value conflicts in public AI governance: a procedural justice framework. *Gov Inf Q*. 2025;42(2):102033.
<https://doi.org/10.1016/j.giq.2025.102033>.

Papagiannidis E, Mikalef P, Conboy K. Responsible artificial intelligence governance: a review and research framework. *J Strateg Inf Syst*. 2025;34(2):101885.
<https://doi.org/10.1016/j.jsis.2024.101885>.

Birkstedt T, Minkkinen M, Tandon A, Mäntymäki M. AI governance: themes, knowledge gaps and future agendas. *Internet Res*. 2023;33(7):133-67.
<https://doi.org/10.1108/INTR-01-2022-0042>.

Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial intelligence. *Gov Inf Q*. 2020;37(3):101493.
<https://doi.org/10.1016/j.giq.2020.101493>.

Balasubramaniam N, Kauppinen M, Rannisto A, Hiekkanen K, Kujala S. Transparency and explainability of AI systems: from ethical guidelines to requirements. *Inf Softw Technol*. 2023;159:107197.
<https://doi.org/10.1016/j.infsof.2023.107197>.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics*. 2020;26(4):2141-68.
<https://doi.org/10.1007/s11948-019-00165-5>.

Floridi L, Cowls J, King TC, Taddeo M. How to design AI for social good: seven essential factors. *Sci Eng Ethics*. 2020;26(3):1771-96.
<https://doi.org/10.1007/s11948-020-00213-5>.

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf Fusion*. 2024;106:102301.
<https://doi.org/10.1016/j.inffus.2024.102301>.

Haque AKMB, Islam AKMN, Mikalef P. Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. *Technol Forecast Soc Change*. 2023;186:122120.
<https://doi.org/10.1016/j.techfore.2022.122120>.

Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495-504.
<https://doi.org/10.1080/10447318.2020.1741118>.