

REVIEW

Open access

Making AI Decisions Answerable: A Governance Framework for Transparency Duties, Auditable Evidence, Contestability, Human Review, and Meaningful Redress

Michael Osei¹, Ingrid Larsen^{2*}, Rajesh Menon³

Abstract

Artificial intelligence decisions are often described as explainable when a system can produce reasons, feature contributions, counterfactuals, or other interpretive outputs. Such outputs, however, do not by themselves make an institutional decision answerable. This article develops a conceptual explainable AI governance framework that treats explainability as an organizational obligation connected to transparency duties, system documentation, evidence preservation, auditability, responsibility allocation, contestability, human review, and redress. The framework distinguishes technical explanation from the governance capacities required to identify who must justify a decision, what information must be recorded, how evidence can be examined, how affected stakeholders may challenge an outcome, and whether an organization can correct errors or remediate consequences. Transparency and documentation are organized around system, model, data, dataset, and decision records, together with data, model, and decision provenance. Auditability is connected to traceability, lifecycle records, risk documentation, monitoring, incident response, and clearly allocated organizational responsibility. Contestability is connected to decision justification, appeal and complaint pathways, competent and sufficiently independent human review, escalation, correction, recourse, and redress capacity. The framework also identifies governance gaps that arise when these functions are separated across the AI lifecycle, including documentation, evidence-trail, accountability, human-review, monitoring, and remediation gaps. Its principal contribution is an answerability-oriented governance architecture that links explanation to institutional evidence and response mechanisms while preserving explicit boundaries: transparency does not independently establish accountability, auditability does not prove responsibility, human involvement does not guarantee meaningful review, and a conceptual governance framework does not establish legal compliance, procedural fairness, ethical sufficiency, or deployment readiness.

Keywords Explainable AI, Accountability, Transparency, Auditability, AI governance, Contestability

*Correspondence:

Ingrid Larsen

ingrid.larsen@bi.no

¹ Department of AI Governance and Accountability, University of Ghana Business School, Accra, Ghana

² Department of Technology Management, BI Norwegian Business School, Oslo, Norway

³ Department of Information Systems, Indian Institute of Management Ahmedabad, Ahmedabad, India

Introduction

Explainable artificial intelligence has become central to debates about responsible intelligent systems, yet explanation availability and institutional answerability remain distinct. Explanation can support responsible AI, but explanation is socially situated and transparency alone does not establish accountability [1-3]. An explanation may describe model

behaviour while leaving unresolved who selected the model, authorized its use, interpreted its output, accepted its risks, or retained authority to correct the resulting decision.

Answerability therefore concerns relationships among reasons, evidence, responsibility, and institutional response. Algorithmic accountability requires actors to justify decisions and remain responsible for the consequences of designing, deploying, or relying on automated systems [4]. This conception extends beyond technical interpretability. It asks whether relevant actors can reconstruct a decision, identify its organizational owner, disclose appropriate information, respond to challenge, and modify either the decision or the underlying governance process.

AI decisions frequently pass through distributed arrangements involving data providers, developers, model owners, deployers, operators, reviewers, and organizational decision-makers. Fragmentation can separate explanation production from decision authority and separate audit evidence from corrective capacity. A technically accurate explanation may consequently have limited governance value when documentation is incomplete, responsibility is diffuse, review is ceremonial, or no pathway connects a successful challenge to correction or remediation.

This article proposes an original conceptual framework for making AI decisions answerable through connected governance functions. It addresses explainability as a governance obligation; transparency and documentation duties; auditable evidence and organizational accountability; contestability, appeals, human review, and redress; lifecycle governance gaps; and implementation risks. The framework is intended to clarify governance responsibilities and evidence requirements. It is not a legal standard, compliance model, audit protocol, certification scheme, or proof that a system is ethically sufficient or ready for deployment.

Explainability as a Governance Obligation

Explainability should first be distinguished from interpretability. Interpretability concerns the extent to which relevant actors can understand a model or its decision logic directly, whereas explainability may include post-hoc accounts of an otherwise opaque system. In consequential applications, substituting explanations for feasible interpretable models can obscure avoidable opacity and weaken scrutiny [5]. Governance should therefore require an explicit justification for the selected level and form of opacity rather than assuming that any explanatory output resolves the issue.

Explanation duties are also stakeholder-dependent. Developers may require diagnostic information, operators may need decision boundaries and uncertainty, reviewers may require traceable evidence, and affected persons may need accessible reasons and challenge-relevant information. Because explanation objectives vary across stakeholders, governance must specify who requires an explanation, why it is required, what decision it concerns, and what action the recipient can reasonably take after receiving it [6]. A single explanation format cannot satisfy all technical, organizational, and participatory purposes.

Human-centered evaluation remains necessary because explanation quality cannot be inferred solely from fidelity or technical sophistication. Perceived explainability and causability may influence trust and acceptance, but positive perception does not demonstrate that reliance is justified or that accountability mechanisms are effective [7]. Explanations should therefore be evaluated for accessibility, usefulness, completeness, uncertainty communication, and alignment with the recipient's practical question, while avoiding the assumption that increased trust is inherently desirable.

Within the proposed framework, explainability becomes a governance obligation when an organization defines explanation duties, allocates responsibility for producing and maintaining them, links explanations to contemporaneous evidence, and preserves pathways for review and correction. Transparency requirements must remain contextual because disclosure needs differ across decision settings and stakeholder roles [8]. This obligation is bounded: transparency cannot independently establish accountability, explanation does not guarantee contestability, and governance terminology cannot establish legal compliance without domain-specific institutional and legal interpretation.

Transparency or documentation requirement	Core answerability question	Responsible actor or function	Required evidence or practice	Governance function	Potential output	Failure or misuse risk
Transparency duty	What information is required, by whom, and for what purpose?	System deployer and governance function	Stakeholder-specific disclosure map	Enables informed scrutiny	Transparency statement	Symbolic or excessive disclosure
Explanation duty	What decision-relevant reasons must be communicated?	Decision owner and explanation designer	Accessible rationale with limits and uncertainty	Supports understanding and challenge	Explanation notice	Generic or misleading reasons

System documentation	How is the sociotechnical system organized?	System owner	Actors, architecture, dependencies, workflow, purpose	Defines governance scope	System dossier	Model-only documentation
Model documentation	Which model was used and under what constraints?	Model owner	Version, purpose, validation, limits, changes	Supports technical review	Model record	Selective performance reporting
Data documentation	What data shaped system behaviour?	Data owner or steward	Sources, transformations, exclusions, quality limits	Supports data scrutiny	Data record	Missing collection context
Dataset documentation	How was a dataset created, intended, and maintained?	Dataset creator or steward	Composition, collection, uses, limits, maintenance	Preserves dataset provenance	Dataset dossier	Template completion without substance
Decision documentation	How was the specific outcome produced?	Decision owner and operator	Inputs, output, thresholds, human actions, overrides	Enables reconstruction	Decision file	Omitted workflow stages
Decision record	Who adopted the outcome and when?	Organizational decision owner	Outcome, timestamp, actor, review status	Fixes chronology and responsibility	Decision record	Unclear institutional ownership
Evidence trail	Can records be connected across stages?	Records and governance functions	Linked data, model, decision, review, and incident records	Enables audit and contestation	Evidence package	Fragmented repositories
Data provenance	Where did data originate and how were they changed?	Data steward	Source and transformation lineage	Supports contextual assessment	Data lineage record	Unknown origin or alteration
Model provenance	Which model configuration produced the output?	Model owner or MLOps function	Version, configuration, training state, dependencies	Supports reproducibility	Model lineage record	Version mismatch
Decision provenance	How did model output become an institutional decision?	System deployer and decision owner	Workflow, thresholds, external inputs, human judgment	Connects computation to action	Decision-provenance map	Treating output as the complete decision

Documentation quality	Are records accurate, current, complete, and usable?	Records-assurance function	Quality criteria and periodic review	Prevents documentation theatre	Quality assessment	Stale or inaccessible records
Explanation accessibility	Can the intended recipient access and understand it?	Communication and accessibility functions	Plain language, suitable format, language support	Enables participation	Accessible explanation	Technical exclusion
Explanation usefulness	Does it answer the recipient's practical question?	Explanation designer and domain owner	Human-centered usefulness evaluation	Reduces irrelevant disclosure	Usefulness assessment	Detailed but unusable output
Explanation completeness	Are relevant factors, human inputs, and limitations disclosed?	Decision owner	Completeness review	Prevents selective justification	Completeness statement	Cherry-picked reasons
Explanation uncertainty	What uncertainty affects the output or explanation?	Model owner and decision owner	Calibrated uncertainty and limitation statement	Supports cautious interpretation	Uncertainty notice	False certainty
No transparency-sufficiency boundary	What remains unproven after disclosure?	Governance function	Explicit limitation statement	Prevents transparency from being equated with accountability	Sufficiency disclaimer	Transparency washing

Auditability and Organizational Accountability

Auditable answerability requires system-oriented audit processes, explicit organizational responsibility, and governance practices that preserve reviewable evidence [13-15]. Auditability is the capacity for an authorized and competent reviewer to examine relevant records, assumptions, decisions, and controls. An audit trail contributes chronological evidence, but auditability also requires meaningful access, interpretable records, appropriate scope, and the ability to connect findings to actors who possess authority to respond.

Organizational AI governance remains fragmented across technical controls, ethical principles, risk functions, and managerial structures [16]. The proposed framework addresses this fragmentation through traceability across the lifecycle. Data provenance should connect to model provenance; model provenance should connect to deployment configurations; and decision provenance should connect model outputs to human judgments, institutional rules, and final outcomes. Lifecycle documentation must also record risk assessments, changes, monitoring findings, incidents, and corrective actions.

Responsibility mapping should identify the substantive authority of each actor. Model owners are responsible for technical stewardship, validation records, version control, and known limitations. System deployers are responsible for use context, workflow configuration, monitoring, and the consequences of integrating outputs into decisions. Operators require defined authority and decision boundaries. Auditors, where supported, require sufficient independence, competence, scope, and evidence access. A governance board or equivalent function may coordinate unresolved issues, but its existence does not prove organizational accountability [14].

Monitoring extends auditability beyond periodic examination. It should capture model and data drift, changes in use, recurring complaints, review outcomes, unusual overrides, and incidents rather than focusing only on predictive performance. Incident reporting should preserve evidence and trigger authorized investigation, containment, communication, and remediation. These mechanisms may support institutional accountability, yet explicit boundaries remain necessary: audit logs do not prove responsibility, auditability does not guarantee corrective action, and formal accountability maps do not demonstrate that an organization has the capacity or willingness to respond [13].

Contestability, Appeals, and Human Redress

Contestability depends on lifecycle responsiveness, capable human oversight, and governance arrangements that prevent responsibility from being abdicated [17–19]. A contestable system allows an affected person, decision subject, end user, or authorized representative to question the decision, its evidence, its justification, or the process by which it was reached. This requires more than an explanation interface: it requires an accessible challenge channel, preserved records, review authority, and an institutional obligation to respond.

Appeal, complaint, recourse, and redress should remain conceptually separate. An appeal requests reconsideration of a decision; a complaint may address broader procedural or institutional conduct; recourse concerns feasible actions that may alter a future outcome; and redress concerns correction or remediation of adverse consequences. Algorithmic recourse is meaningful only when recommended actions are feasible, consequential, and sufficiently stable rather than merely contrastive descriptions of how an output might differ [20].

An answerable decision pathway begins with an identifiable AI-supported outcome and its institutional context. Transparency duties and documentation create the basis for an evidence trail; decision justification connects reasons to that evidence; contestability initiates challenge; and meaningful human review examines the challenged decision with appropriate competence, independence, time, and authority. Escalation should connect unresolved disputes to correction, remediation, and governance updating. **Figure 1** illustrates the answerability pathway through which AI decisions can move from explanation and evidence trails to contestability, human review, escalation, correction, and meaningful redress.

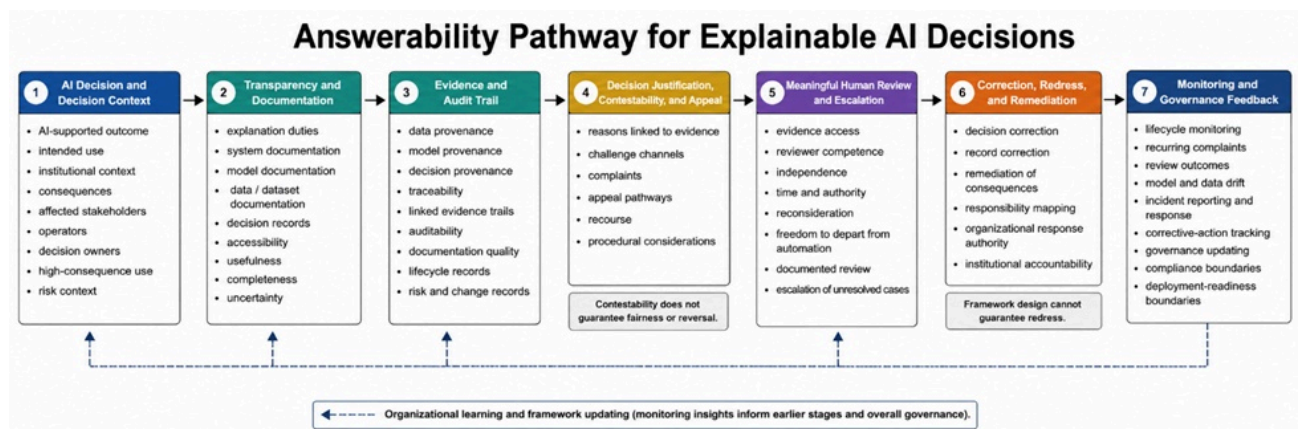


Figure 1. Answerability Pathway for Explainable AI Decisions

Alt-text description:

A conceptual pathway diagram showing AI decision output connected to transparency duties, documentation, evidence trails, decision justification, contestability, appeal, human review, escalation, correction, redress, monitoring, and governance feedback.

Human review should not be treated as meaningful merely because a person appears in the workflow. Review quality depends on access to relevant evidence, freedom to depart from the automated recommendation, competence to evaluate technical and domain considerations, and authority to trigger correction or escalation. Contestability likewise does not guarantee procedural fairness, appeal does not guarantee reversal, and redress cannot be claimed where an organization lacks practical correction or remediation capacity.

Governance Gaps Across the AI Lifecycle

Governance gaps often emerge because organizations adopt broad principles without establishing the mechanisms required to apply them. Ethical principles can orient decision-making, but they do not independently allocate responsibility, generate evidence, create review authority, or ensure corrective action [21]. A principle-to-practice gap therefore arises when responsible-AI language is not connected to lifecycle duties and institutional capacity.

Documentation, assessment, auditing, stakeholder engagement, and monitoring tools operate at different lifecycle stages and serve different purposes [22]. A documentation gap may originate during data collection, whereas an evidence-trail gap may become visible only when an individual decision is challenged. Auditability gaps appear when records exist but cannot be accessed or interpreted, and decision-justification gaps arise when reasons are reconstructed after the outcome rather than recorded contemporaneously.

High-level convergence around transparency, fairness, responsibility, and related principles can conceal substantial differences in interpretation and implementation [23]. Symbolic transparency may disclose general system information while omitting evidence needed to understand a particular decision. Similar gaps arise when responsibility is described collectively but no actor has sufficient authority to investigate, respond, or correct.

Responsible-AI governance requires structural, procedural, and relational mechanisms across the lifecycle [24]. Structural mechanisms allocate authority; procedural mechanisms establish documentation, review, and escalation; relational mechanisms coordinate stakeholders and organizational functions. When these mechanisms are disconnected, responsibility ambiguity, weak human review, contestability without remedy, monitoring failure, and delayed incident response may persist despite extensive governance language.

Public-governance research further demonstrates the importance of context, impact, multidisciplinary assessment, and continuing evaluation [25]. High-consequence systems may require more demanding evidence and review arrangements, but the framework does not prescribe universal thresholds. Model drift, data drift, changing use contexts, unresolved complaints, and incidents should instead trigger context-specific reassessment. Table 2 maps auditability, organizational accountability, contestability, appeals, human redress, and lifecycle governance gaps across AI decision systems.

Table 2. Auditability, Organizational Accountability, Contestability, Appeals, Human Redress, and Lifecycle Governance Gaps in Explainable AI Systems: Audit Trails, Traceability, Responsibility Mapping, Human Review, Review Independence, Escalation, Decision Justification, Redress, Monitoring, Incident Response, and Governance Boundaries

Governance requirement or gap	Lifecycle or decision stage	Core governance question	Responsible actor or function	Required evidence or procedure	Potential output or consequence	Failure or responsibility gap
-------------------------------	-----------------------------	--------------------------	-------------------------------	--------------------------------	---------------------------------	-------------------------------

Auditability	Validation and operation	Can a competent reviewer examine relevant evidence?	System owner and audit function	Defined scope, access, criteria, and evidence	Audit-ready evidence set	Records exist but cannot support review
Audit trail	Development to retirement	What actions and changes occurred?	Technical and records functions	Contextualized logs and change records	Reconstructed chronology	Logs lack meaning or ownership
Traceability	Entire lifecycle	Can data, model, human action, and outcome be linked?	Data, model, deployment, and governance owners	End-to-end lineage	Traceability map	Broken evidence connections
Responsibility mapping	Procurement and deployment	Who controls each material decision?	Governance function	Role, authority, and escalation map	Responsibility matrix	Shared responsibility becomes no responsibility
Human-review gap	Decision and appeal	Can reviewers exercise informed judgment?	Review owner	Evidence access, competence, authority, reasoning record	Substantive reconsideration or confirmation	Automation bias or rubber stamping
Review independence	Appeal	Is the reviewer sufficiently separated from the original decision?	Appeal function	Conflict assessment and alternative escalation	Independent review record	Self-review or managerial pressure
Contestability gap	Post-decision	Can stakeholders challenge the outcome or process?	System deployer and complaint function	Accessible challenge channel and decision evidence	Contestation record	Challenge without institutional response
Appeal and complaint pathways	Post-decision	Are reconsideration and broader complaints distinguished?	Review and complaint functions	Clear intake, routing, timing, and authority	Appeal or complaint decision	Confused or duplicative procedures
Decision-justification gap	Decision stage	Are reasons linked to contemporaneous evidence?	Decision owner	Reason statement tied to decision provenance	Justification record	Post-hoc rationalization

Redress gap	Correction and remediation	Can valid findings alter records, decisions, or consequences?	Authorized institutional function	Correction power, remedy options, follow-through	Correction or remediation	Symbolic acknowledgement
Monitoring gap	Post-deployment	Are changes in performance, use, impact, and complaints detected?	Monitoring owner	Technical, procedural, incident, and stakeholder indicators	Monitoring report	Accuracy-only monitoring
Drift gap	Post-deployment	Have model or data assumptions changed?	Model and data owners	Baselines, thresholds, contextual review	Drift-review record	Undetected degradation
Incident-response gap	Post-deployment	Can failures be contained and investigated promptly?	Incident-response function	Reporting criteria, authority, preserved evidence	Response and remediation record	Delayed escalation or repeated harm
Accountability gap	Organization-wide	Who must justify, answer, and correct?	System-deploying organization	Institutional authority and response duties	Accountability map	Outsourced or diffused responsibility
No compliance claim boundary	Governance review	Is the framework being treated as legal proof?	Governance and legal functions where applicable	Explicit claim limitation	Compliance-boundary statement	False assurance
No deployment-readiness boundary	Approval and change review	Is checklist completion being mistaken for readiness?	Decision authority	Contextual validation and capacity assessment	Readiness limitation statement	Premature deployment

Proposed Policy and Organizational Framework

The proposed framework begins with AI use-case and decision-context definition. Organizations should identify the institutional decision, intended role of the AI system, foreseeable consequences, affected stakeholders, and limits of automated influence. Ethical principles such as autonomy, beneficence, non-maleficence, justice, and explicability can inform this stage, but they must be translated into context-specific governance questions rather than treated as sufficient operational rules [26].

The second stage maps stakeholders, responsibilities, and decision rights. Human-centered governance should combine reliable technical design with organizational controls, human authority, monitoring, and channels for reporting failure [27]. The framework therefore distinguishes developers, data owners, model owners, system deployers, operators, reviewers,

The third stage allocates transparency duties and defines documentation, evidence-trail, auditability, traceability, human-review, contestability, appeal, complaint, and redress requirements. Stakeholder participation should inform these requirements because system values and acceptable trade-offs cannot be determined only by technical actors [28]. Participation is not a transfer of responsibility to affected stakeholders; the organization retains responsibility for interpreting feedback and acting on material findings.

The final stage establishes review and updating. Values embedded during design require continuing verification because implementation, use, and context can alter whether those values are realized [30]. Audit findings, appeals, complaints, incidents, drift, and stakeholder feedback should therefore update documentation, controls, and responsibility arrangements. **Table 3** presents the proposed policy and organizational framework for making AI decisions answerable through transparency duties, auditable evidence, contestability, human review, and meaningful redress.

[illegible]

Evidence-trail design	Link data, model, decision, review, and incident records	Documentation architecture	Enables reconstruction	Evidence specification	Traceability testing	Fragmented evidence	Challenges
Auditability and traceability design	Enable competent examination	Evidence trail and review objectives	Supports scrutiny	Auditability plan	Scope, access, competence, independence	Audit theatre	Control actions
Human-review pathway	Define substantive reconsideration	Risk, reviewer role, evidence needs	Adds contextual judgment	Review protocol	Competence, authority, workload, independence	Rubber stamping	Control actions
Contestability pathway	Enable structured challenge	Decision evidence and access needs	Supports answerability	Contestation procedure	Burden and accessibility evaluation	Challenge without response	Control actions
Appeal and complaint design	Route reconsideration and process concerns	Institutional authority	Separates review functions	Appeal and complaint map	Timeliness and independence review	Confused pathways	Control actions
Redress and remediation design	Connect valid findings to correction and learning	Remedy authority and root-cause evidence	Converts review into action	Redress plan	Capacity and follow-through review	Symbolic remedy	Control actions
Decision-justification design	Link reasons to evidence	Decision factors and human inputs	Creates answerable rationale	Justification record	Completeness and fidelity review	Post-hoc rationalization	Control actions
Risk and impact review	Identify, assign, and monitor risks	Technical, organizational, and stakeholder evidence	Coordinates risk decisions	Risk record	Independent challenge where appropriate	Checklist assessment	Monitor incidents
Lifecycle monitoring and drift review	Detect technical and contextual change	Baselines, indicators, complaints	Triggers reassessment	Monitoring and drift records	Periodic governance review	Unseen degradation	Respond to incidents
Incident reporting and response	Detect, contain, investigate, and correct failures	Incident taxonomy and authority map	Supports failure management	Incident record	Simulation and post-incident review	Delayed action	Learn from incidents

Organizational accountability mapping	Connect duties to authority and response capacity	Role map and governance structures	Establishes institutional answerability	Accountability architecture	Senior governance review	Responsibility without authority	A
Governance review and updating	Revise framework as evidence changes	Audits, appeals, incidents, drift, feedback	Enables organizational learning	Updated governance framework	Documented change approval	Static governance	Co
Claim and deployment boundaries	Prevent legal, ethical, and operational overstatement	All framework evidence and limitations	Controls interpretation	Boundary statement	Domain-specific validation	Governance washing	E

Figure 2 presents the proposed policy and organizational framework for answerable AI decisions, integrating transparency duties, documentation, auditable evidence, contestability, human review, redress, lifecycle monitoring, and governance updating.

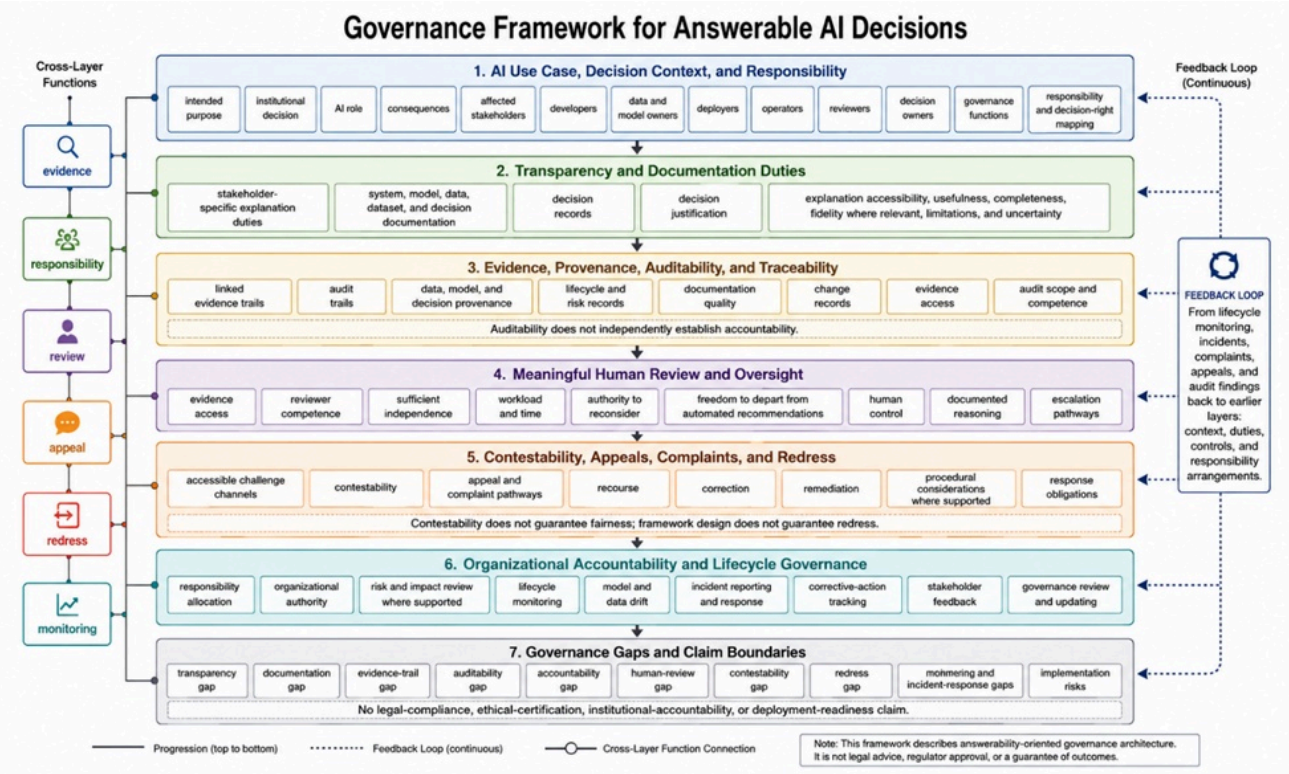


Figure 2. Governance Framework for Answerable AI Decisions

Alt-text description:

A conceptual governance framework showing AI use-case definition, responsibility mapping, transparency duties, documentation, evidence trails, auditability, traceability, human review, contestability, appeals, redress, monitoring, incident response, organizational accountability, and governance updating.

Implementation Risks and Recommendations

The first implementation risk is governance symbolism. Transparency statements, documentation templates, ethics principles, and committees may produce visible activity without usable evidence or corrective authority. Transparency should therefore be evaluated against concrete stakeholder questions, while documentation should be tested for accuracy, currency, accessibility, and relevance. Organizations should avoid treating disclosure volume as a proxy for answerability, particularly because visibility alone cannot resolve institutional accountability gaps [3].

A second risk is evidence fragmentation. Audit logs may record technical events while omitting decision context, human judgment, responsibility, or downstream consequences. Evidence-trail design should therefore connect system, model, data, decision, review, and incident records. Research should examine how such evidence can remain proportionate, privacy-conscious, interpretable, and resistant to retrospective rationalization. Operational records should support governance inquiry rather than become passive archives [11].

A third risk is procedural façade. Human reviewers may defer to automated recommendations; contestability may impose heavy burdens on affected persons; and appeal channels may lack authority to correct outcomes. Reviewer competence, independence, workload, evidence access, and intervention authority should be evaluated directly. Challenge pathways should not require individuals to reproduce information already controlled by the organization, and their effectiveness should be assessed by responsiveness rather than mere availability [18].

A fourth risk concerns organizational capacity and overprescription. The proposed framework requires staff, domain expertise, records management, authority, and correction resources that may be unevenly available. Implementation should therefore be treated as a context-specific research and governance problem rather than a universal checklist. The recommendations are conceptual and do not constitute legal advice, compliance guidance, policy prescription, audit instruction, procurement advice, platform-selection advice, or deployment instruction [24].

Conclusion

This article has proposed a conceptual governance framework for making AI decisions answerable through connected transparency duties, system and decision documentation, auditable evidence, traceability, responsibility mapping, contestability, appeals, meaningful human review, correction, redress, lifecycle monitoring, incident response, and governance updating. Its central claim is that explanation availability becomes institutionally valuable only when explanations are connected to evidence, authorized responsibility, effective challenge, review quality, and practical remediation capacity. Answerable AI should therefore be evaluated through the organization's ability to preserve relevant records, justify decisions, respond to challenge, assign responsibility, correct failures, and learn from monitoring and incidents. The framework does not establish legal compliance, procedural fairness, ethical sufficiency, institutional accountability, or deployment readiness; it offers a structured basis for developing and examining those governance capacities within specific domains.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Aug 2025

Revised: 23 Sep 2025

Accepted: 06 Nov 2025

Published online: 15 March 2026

Rights and permissions

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

References

- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.
- 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>.
- 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>.
- Horneber D, Laumer S. Algorithmic accountability. *Bus Inf Syst Eng*. 2023;65(6):723-30.
<https://doi.org/10.1007/s12599-023-00817-8>.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.
- Meske C, Bunde E, Schneider J, Gersch M. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Inf Syst Manag*. 2022;39(1):53-63.
<https://doi.org/10.1080/10580530.2020.1849465>.
- 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>.
- Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc*. 2019;6(1):2053951719860542.
<https://doi.org/10.1177/2053951719860542>.

Geburu T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. *Commun ACM*. 2021;64(12):86-92.
<https://doi.org/10.1145/3458723>.

Bender EM, Friedman B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans Assoc Comput Linguist*. 2018;6:587-604.
https://doi.org/10.1162/tacl_a_00041.

Williams R, Cloete R, Cobbe J, Cottrill C, Edwards P, Markovic M, et al. From transparency to accountability of intelligent systems: moving beyond aspirations. *Data Policy*. 2022;4:e7.
<https://doi.org/10.1017/dap.2021.37>.

de Fine Licht K, de Fine Licht J. Artificial intelligence, transparency, and public decision-making. *AI Soc*. 2020;35(4):917-26.
<https://doi.org/10.1007/s00146-020-00960-w>.

Mökander J, Floridi L. Ethics-based auditing to develop trustworthy AI. *Minds Mach*. 2021;31(2):323-7.
<https://doi.org/10.1007/s11023-021-09557-8>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics*. 2019;160(4):835-50.
<https://doi.org/10.1007/s10551-018-3921-3>.

Papagiannidis E, Enholm IM, Dremel C, Mikalef P, Krogstie J. Toward AI governance: identifying best practices and potential barriers and outcomes. *Inf Syst Front*. 2023;25(1):123-41.
<https://doi.org/10.1007/s10796-022-10251-y>.

Birkstedt T, Minkinen 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>.

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>.

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>.

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>.

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>.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1(11):501-7.
<https://doi.org/10.1038/s42256-019-0114-4>.

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>.

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>.

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>.

Zuiderwijk A, Chen YC, Salem F. Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. *Gov Inf Q*. 2021;38(3):101577.

<https://doi.org/10.1016/j.giq.2021.101577>.

Floridi L, Cows J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 2018;28(4):689-707.

<https://doi.org/10.1007/s11023-018-9482-5>.

Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans Interact Intell Syst*. 2020;10(4):1-31.

<https://doi.org/10.1145/3419764>.

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>.

Mäntymäki M, Minkinen M, Birkstedt T, Viljanen M. Defining organizational AI governance. *AI Ethics*. 2022;2(4):603-9.

<https://doi.org/10.1007/s43681-022-00143-x>.

van de Poel I. Embedding values in artificial intelligence (AI) systems. *Minds Mach*. 2020;30(3):385-409.

<https://doi.org/10.1007/s11023-020-09537-4>.