

ORIGINAL RESEARCH

Open access

The Hidden Cost of Unexplained AI: A Conceptual Model of Explainability Debt, Documentation Gaps, Interface Misalignment, Stakeholder Exclusion, and Accumulating Governance Risk

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

Abstract

Artificial intelligence systems may remain operationally usable while carrying unresolved explanation-related deficiencies that are difficult to observe during initial development. These deficiencies can emerge later as opaque model behavior, uncertain data provenance, incomplete documentation, broken traceability, unsuitable explanation interfaces, excluded stakeholder perspectives, weak contestability, and ambiguous accountability. This article introduces explainability debt as a conceptual lifecycle risk arising when explanation design, documentation, traceability, stakeholder engagement, interface alignment, and governance review are deferred or inadequately maintained. The proposed approach adapts technical-debt reasoning without treating explainability debt as a measurable financial liability or validated risk score. It distinguishes model and data opacity from the organizational deficiencies that prevent opacity from being investigated, communicated, challenged, and corrected. Documentation and traceability deficits are interpreted as missing or deteriorating evidence relationships across datasets, model versions, explanations, decisions, and responsible actors. Interface misalignment and stakeholder exclusion are treated as mechanisms through which technical opacity can become misunderstanding, inappropriate reliance, limited participation, and weak reviewability. Unresolved deficiencies may subsequently contribute to accountability gaps, operational uncertainty, decision risk, and broader governance risk. The article's main contribution is a conceptual model connecting deferred explainability work with accumulating lifecycle costs and feedback failures across development, deployment, monitoring, incident learning, correction, and model updating. The model supports structured reasoning about hidden explainability obligations but does not measure debt, prove trustworthiness, guarantee human understanding, establish legal or ethical sufficiency, certify governance, or determine deployment readiness.

Keywords Explainable AI, Trustworthy AI, Traceability, Explainability debt, Documentation, Stakeholder participation

*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 encompasses diverse technical methods, interpretive objectives, and human communication requirements. Explanations may describe features, representations, rules, examples, counterfactuals, or

local decision rationales, yet their usefulness depends on the question being asked and the person receiving them. In high-stakes settings, a plausible post-hoc explanation cannot automatically remove the risks created by an opaque model [1-3].

Explainability is also not equivalent to trustworthiness. Trustworthy AI depends on several interacting conditions, including technical reliability, transparency, human oversight, organizational responsibility, and appropriate governance. An AI system may therefore produce an explanation while remaining poorly documented, difficult to audit, unsuitable for its users, or weakly governed [4].

This article uses the term explainability debt to describe unresolved explanation-related obligations that accumulate across the AI lifecycle. The concept includes more than the absence of an explanation output. It covers deferred explanation requirements, model and data opacity, incomplete documentation, broken evidence trails, inaccessible interfaces, excluded stakeholders, unclear responsibilities, and failures to update explanatory resources when systems change [1].

The objective is to propose an original conceptual model showing how these deficits can create hidden lifecycle costs. The model connects technical opacity with human interpretation and organizational governance, while preserving important boundaries: it does not assign numerical debt values, establish legal accountability, prove ethical adequacy, or authorize deployment. Instead, it structures reasoning about when explainability should be treated as a continuing lifecycle responsibility rather than a late-stage add-on [4].

Explainability Debt as a Lifecycle Risk

Technical debt describes the future burden created when expedient or incomplete work is accepted in the present. Debt can remain latent until maintenance, redesign, integration, or correction becomes necessary. Its consequences depend not only on the original shortcut but also on whether responsibility, knowledge, and resources exist to identify and address it [5, 6].

AI-enabled systems extend this problem because model behavior depends on changing data, pipelines, architectures, operating contexts, and monitoring practices. AI-related debt can affect understandability, maintainability, architecture, and quality, while newer development environments introduce additional dependencies across DevOps, MLOps, and generative-AI lifecycles [7, 8].

Explainability debt is proposed here as a narrower but interconnected form of lifecycle debt. It accumulates when organizations postpone explanation design, documentation, stakeholder engagement, interface evaluation, or governance review. A system may initially function despite these omissions, but later changes can expose missing evidence, stale explanations, unclear ownership, or interfaces that no longer match the decision context [7].

The concept includes latent debt, accumulated debt, model-update risk, monitoring gaps, lifecycle-feedback gaps, incident-learning gaps, and correction gaps. These categories should be interpreted qualitatively until validated measurement approaches exist. Identifying a potential debt source does not demonstrate its severity, predict failure, eliminate residual risk, or determine whether an AI system is ready for deployment [8].

Accumulating Model and Data Opacity

Model opacity concerns the difficulty of understanding how an AI system transforms inputs into outputs. Black-box behavior can arise from complex structures, inaccessible internal representations, nonlinear feature interactions, or decision logic that cannot be reconstructed meaningfully. Because different opacity sources require different explanatory approaches, no single technique can address every model, stakeholder, or decision context [9].

Deep learning introduces several distinguishable forms of opacity. Feature processing may obscure how variables are transformed; internal representations may lack clear semantic meaning; and the relationship between learned representations and final outputs may be difficult to communicate. Post-hoc methods can reveal patterns, but their outputs require careful evaluation and should not be treated as complete replicas of model reasoning [10].

Opacity also has an evaluative dimension. Explainability research contains multiple definitions, objectives, and assessment approaches, with no universally sufficient measure of explanation quality. Technical fidelity, human comprehensibility, usefulness, and decision relevance may produce different conclusions about the same explanation [11].

Data opacity develops when the origin, composition, quality, labeling, transformation, or contextual limitations of data are unclear. Provenance gaps weaken the ability to explain where evidence came from, while undocumented labeling and feature construction can conceal assumptions embedded before model training. Model updates can intensify these problems when new data or retraining alters behavior without corresponding changes to explanations and records [12].

Opacity therefore accumulates through interactions among model complexity, black-box behavior, opaque features and representations, undocumented updates, uncertain data provenance, quality limitations, and weak lineage. Recognizing these sources can guide later documentation and review, but classification alone does not resolve opacity or guarantee faithful explanation. **Table 1** summarises how explainability debt accumulates through lifecycle risks, model opacity, data opacity, deferred explanation design, and deferred governance review.

Table 1. Explainability Debt as a Lifecycle Risk in AI Systems: Deferred Explanation Design, Technical Debt, Machine Learning Technical Debt, Model Opacity, Data Opacity, Opaque Feature Processing, Opaque Representations, Opaque Model Updates, Data Provenance Gaps, Data Quality Gaps, Documentation Gaps, Monitoring Gaps, Lifecycle Feedback Gaps, and Debt Boundaries

Debt source or opacity source	Core lifecycle question	How the debt or opacity appears	Potential hidden cost	Affected lifecycle stage or stakeholder	Detection, documentation, or review need	Boundary statement
Explainability debt	Which explanation obligations remain unresolved?	Explanation, evidence, interface, and review work is deferred	Later redesign, weak reviewability, and governance escalation	Entire lifecycle	Periodic cross-functional review	The concept does not independently measure debt
Technical debt	Which present compromises create future maintenance burdens?	Short-term choices persist beyond their intended period	Rework, fragility, and reduced maintainability	Design, development, and maintenance	Technical-debt identification and ownership	Not every technical defect is explainability debt
Machine-learning technical debt	Which data- and model-dependent dependencies create later burdens?	Pipelines, data, features, and models become tightly coupled	Difficult updating and unstable behavior	Development, deployment, and updating	Lifecycle and dependency review	Broader than explainability debt

Deferred explanation design	Were explanation needs defined too late?	Explanations are added after core design choices are fixed	Incompatible methods and expensive redesign	Requirements and design teams	Early explanation-requirement review	Early planning does not guarantee usefulness
Deferred documentation and governance review	Were records, roles, and review gates postponed?	Evidence remains incomplete and responsibility remains unclear	Costly reconstruction and accountability gaps	Development, deployment, and governance bodies	Documentation and responsibility review	Review cannot eliminate residual risk
Latent and accumulated debt	Which unresolved deficits remain hidden or reinforce one another?	Small omissions persist and combine across stages	Disproportionate correction effort	Entire socio-technical system	Integrated gap assessment	Accumulation remains conceptual without validation
Model opacity and complexity	Which internal mechanisms cannot be meaningfully understood?	High-dimensional or nonlinear processing obscures behavior	Oversimplified or misleading explanations	Developers, reviewers, operators, and decision subjects	Architecture, behavior, and limitation records	Identifying opacity does not resolve it
Opaque feature processing, representation, and decision logic	How are features transformed and connected to outputs?	Internal states and interactions lack accessible meaning	Incorrect attribution or false interpretability	Developers and domain experts	Feature and representation analysis	Post-hoc accounts may remain incomplete
Opaque model updates	What changed between versions and why?	Retraining changes behavior without synchronized explanations	Stale explanations and inappropriate reliance	Maintainers, operators, and reviewers	Version-linked change and explanation records	Updated records do not prove valid explanations
Data opacity and provenance gaps	Where did data originate and under what conditions?	Sources, collection context, or transformations are missing	Weak evidence trails and misuse	Data subjects, developers, and auditors	Provenance and data-lineage documentation	Provenance does not establish data adequacy
Data quality, labeling, and feature-documentation gaps	Are limitations and human choices visible?	Error, missingness, labeling rules, or transformations are undocumented	Hidden bias and misleading certainty	Domain experts and affected stakeholders	Quality, labeling, and feature records	Documentation does not remove data limitations

Monitoring and lifecycle-feedback gaps	Are emerging explanation failures detected and returned to design?	No process captures misunderstanding, drift, or user concerns	Persistent defects and delayed correction	Deployment and operation	Socio-technical monitoring and feedback review	Monitoring cannot observe every failure
Incident-learning and correction gaps	Do identified failures lead to verified changes?	Findings remain isolated, unassigned, or unresolved	Recurrence and institutional forgetting	Incident response and maintenance	Ownership, corrective action, and effectiveness review	Correction records do not prove risk removal
Debt-measurement and deployment boundaries	Can the model quantify debt or authorize deployment?	Conceptual categories are mistaken for validated scores or approvals	False precision and premature assurance	Governance and deployment decisions	Independent empirical and domain-specific assessment	The model neither measures debt nor establishes readiness

Documentation and Traceability Deficits

Dataset documentation can reduce data opacity by recording motivation, composition, collection conditions, intended uses, and known limitations. Such records help later users understand the context in which data were created, but their value depends on completeness, accessibility, maintenance, and continued connection to downstream models and decisions [13].

Contextual documentation is also relevant to stakeholder inclusion. Data statements can disclose population characteristics, language conditions, annotation assumptions, and generalization boundaries that might otherwise remain invisible. When these conditions are omitted, explanation interfaces may present model outputs without revealing whose experiences were represented, simplified, or excluded during data construction [14].

AI documentation extends beyond dataset records. Model, system, decision, and explanation documentation may be needed to support auditability and accountability. Empirical evidence indicates that documentation practices can be inconsistent, burdensome, or poorly adapted to AI-specific risks, leaving reviewers without sufficient evidence to reconstruct how important choices were made [15].

Documentation approaches also remain fragmented. Different instruments describe datasets, models, systems, impacts, or organizational processes, but these records may not share identifiers, update schedules, ownership rules, or compatible levels of detail. As a result, documentation can appear extensive while still failing to provide an integrated evidence trail across the lifecycle [16].

Explainability debt therefore includes documentation completeness, usability, and maintenance deficits as well as traceability, auditability, evidence-trail, decision-trail, model-lineage, data-lineage, and explanation-lineage gaps. These relationships require versioned links among data, models, explanation methods, interfaces, decisions, and responsible actors. Nevertheless, even complete and traceable records cannot independently prove that an explanation is faithful, a decision is justified, accountability is effective, or governance risk has been eliminated [15].

Interface Misalignment and Stakeholder Exclusion

Explanation interfaces mediate between technical outputs and the people expected to interpret, question, or act on them. Their design should therefore begin with the user's role, decision task, expertise, and explanation objective rather than with the availability of a particular visualization or post-hoc method. Evaluation must similarly examine whether an explanation supports the intended activity for the relevant stakeholder [17].

Interface alignment includes explanation scope, format, length, timing, visualization, and degree of user control. Preferences can differ substantially across contexts: some users need concise local reasons, whereas others require system-level information, uncertainty, limitations, or comparative evidence. Excessive detail can produce explanation overload, while insufficient detail can conceal assumptions needed for informed judgment [18].

Misalignment also occurs when an explanation contradicts the associated recommendation, omits decisive information, or answers a different question from the one confronting the user. Such inconsistencies may reduce acceptance or encourage users to interpret explanation fluency as evidence of model validity. Explanation availability must therefore be distinguished from explanation coherence and decision usefulness [19].

Human understanding cannot be inferred from the presence of an interface. Perceived explainability and causability may influence trust and acceptance, but these responses depend on context, prior knowledge, expectations, and the quality of interaction. Comprehension, appropriate reliance, and the ability to recognize limitations require direct human-centered assessment [20].

Stakeholder exclusion compounds interface debt when affected persons, decision subjects, operators, developers, deployers, domain experts, or governance bodies are absent from requirements and evaluation. Participation can expose conflicting explanation needs and unequal access to recourse, but inclusion alone does not prove fairness, legitimacy, accountability, or adequate understanding. The proposed model therefore treats stakeholder participation as a continuing evaluative requirement rather than a one-time consultation [17].

Governance Consequences of Unresolved Explainability Debt

AI ethics and governance frameworks frequently endorse transparency, responsibility, fairness, and accountability, but agreement at the level of principles does not ensure consistency in interpretation or implementation. When explanation obligations remain abstract, organizations may lack evidence showing how principles influenced system requirements, operational decisions, or corrective actions [21].

Principles alone cannot establish accountable conduct. Institutional mandates, professional responsibilities, organizational capabilities, and mechanisms for scrutiny are also required. Explainability debt can therefore contribute to governance debt when transparency claims are disconnected from authority, evidence, resources, and procedures for intervention [22].

The transition from principles to practice requires methods suited to different lifecycle stages. Documentation, impact analysis, monitoring, stakeholder engagement, and review mechanisms may help operationalize governance, but fragmented or deferred implementation can leave significant gaps between stated commitments and actual system control [23].

Ethics-based auditing can support review by examining evidence, processes, impacts, and organizational controls. Nevertheless, an audit remains constrained by its scope, criteria, evidence access, reviewer independence, and timing. Auditability is therefore an important governance function but not proof that an AI system is trustworthy, accountable, legally sufficient, or free from residual risk [24].

Unresolved explainability debt may appear as contestability gaps, weak reviewability, unclear responsibility mapping, ineffective role allocation, traceability failure, operational uncertainty, decision risk, or reputational exposure where such consequences are supported by context. Monitoring, incident learning, and correction mechanisms are necessary to prevent known deficiencies from persisting, but they cannot eliminate governance risk or guarantee accountable conduct [24]. **Table 2** maps documentation deficits, traceability gaps, interface misalignment, stakeholder exclusion, and governance consequences that can arise when explainability debt remains unresolved.

Table 2. Documentation, Traceability, Interface, Stakeholder, and Governance Consequences of Explainability Debt: Model Documentation Gaps, Data Documentation Gaps, Decision Documentation Gaps, Evidence Trail Gaps, Explanation Interfaces, Explanation Overload, Stakeholder Exclusion, Contestability Gaps, Reviewability Gaps, Accountability Gaps, Operational Risk, Decision Risk, and Governance-Risk Boundaries

Debt consequence or requirement	Core governance or human-centered question	Required evidence, practice, or participation condition	Explainability or governance function	Potential output	Failure or residual risk	Maintenance, evaluation, or accountability need
Model, data, and system documentation gaps	Are purpose, assumptions, data context, architecture, and limitations recorded?	Versioned and accessible records	Contextualizes models and explanations	Reviewable documentation set	Incomplete or misleading evidence	Assigned ownership and update triggers
Decision and explanation documentation gaps	Can a decision and its explanation be reconstructed?	Inputs, outputs, explanation method, overrides, timing, and roles	Supports review and contestability	Decision and explanation trail	Unreviewable consequential decisions	Retention, access, and privacy governance
Documentation completeness, usability, and maintenance deficits	Are records sufficient, understandable, and current?	Completeness criteria, user testing, and change-linked updates	Makes evidence usable across the lifecycle	Documentation-gap findings	Documentation theatre or stale evidence	Periodic quality and currency review
Traceability and evidence-trail gaps	Can claims be linked to models, data, processes, and responsible actions?	Persistent identifiers and connected artifacts	Supports reconstruction and audit	Traceability map	Unsupported explanation claims	Independent reconstruction review
Model, data, and explanation-lineage gaps	Can versions and dependencies be followed across changes?	Linked version histories	Prevents stale explanation use	Lineage record	Version mismatch and irreproducibility	Release and change governance

Explanation interface and accessibility requirements	Can intended stakeholders access and use the explanation?	Task-sensitive, accessible presentation	Communicates relevant evidence and limitations	Evaluated explanation interface	Exclusion or misinterpretation	Accessibility, usability, and comprehension testing
Explanation overload, under-specification, and mismatch	Does the explanation provide suitable depth and answer the relevant question?	Role-, expertise-, and case-sensitive content	Supports cognitive accessibility	Interface-alignment findings	Confusion, false simplicity, or inappropriate reliance	Task-based evaluation
Stakeholder participation and exclusion	Whose needs and harms influence requirements and review?	Meaningful participation and evidence of influence	Introduces situated explanation needs	Stakeholder-gap analysis	Tokenism or omitted perspectives	Participation-process and outcome evaluation
Contestability and reviewability gaps	Can an affected party challenge an outcome and can a reviewer assess it?	Accessible reasons, evidence, records, and review procedures	Supports scrutiny and recourse	Contestability and reviewability assessment	Formal channels without meaningful challenge	Independent procedural evaluation
Accountability, responsibility, and role-allocation gaps	Who has authority and responsibility for explanation-related obligations?	Role mapping, escalation paths, competence, and resources	Connects evidence to accountable action	Responsibility map	Blame shifting or unresolved findings	Review of actual authority and performance
Operational, decision, and reputational risk	How may unresolved deficits affect use and organizational credibility?	Contextual risk evidence and monitored incidents	Connects debt with potential consequences	Qualitative risk narrative	Overstated causality or omitted interactions	Domain-specific risk assessment
Governance-risk and accountability boundaries	Does the model eliminate risk or guarantee accountability?	Independent technical, human, institutional, and domain review	Preserves residual-risk reasoning	Explicit boundary statement	False assurance from framework use	Continuing governance scrutiny

Proposed Explainability Debt Model

The proposed model begins with use-case and lifecycle-context definition. The system's purpose, decision stakes, operating environment, affected groups, organizational setting, and lifecycle stage establish which explanation obligations

may be relevant. Organizational AI governance research indicates that effective oversight depends on coordinated structures, processes, roles, and relational mechanisms rather than a single transparency intervention [25].

The second component maps stakeholders and decision contexts before defining explanation requirements. Developers, operators, deployers, decision subjects, affected persons, domain experts, and governance bodies may require different evidence and levels of detail. Governance practices are also shaped by organizational capabilities, culture, role clarity, monitoring arrangements, and implementation barriers [26].

The model then examines model opacity, data opacity, documentation gaps, traceability gaps, interface alignment, and stakeholder inclusion as connected but non-equivalent debt pathways. These reviews should begin early and persist through updates because governance affordances, assurance evidence, and explanation resources are more difficult to reconstruct after design commitments and operational dependencies have accumulated [27].

Contestability, reviewability, accountability, and responsibility mapping form the governance-assessment layer. Accountability should not be treated as a single undifferentiated obligation: answerability for explaining conduct and responsibility for actions or outcomes may involve different actors, powers, and evidence requirements. Formal allocation is useful only when roles are supported by authority, competence, resources, and review mechanisms [28].

Finally, the model connects identified deficits with a context-specific debt register, qualitative prioritization, bounded remediation planning, monitoring, incident learning, correction feedback, and model updating. Stakeholders should have agency to request task-relevant information and interact with explanations, but increased trust or perceived transparency does not necessarily improve human–AI team performance [29]. **Table 3** presents the proposed explainability debt model linking lifecycle context, stakeholder mapping, opacity review, documentation, traceability, interface alignment, stakeholder inclusion, governance risk, monitoring, and model updating.

Table 3. Proposed Explainability Debt Model for Trustworthy AI Lifecycle Governance: Lifecycle Context, Stakeholder Mapping, Explanation Requirements, Model Opacity Review, Data Opacity Review, Documentation Gap Review, Traceability Gap Review, Interface Alignment, Stakeholder Inclusion, Contestability, Accountability, Governance Risk, Debt Prioritization, Remediation Planning, Monitoring, Incident Learning, and Deployment Boundaries

Model component	Core purpose	Required input	Explainability-debt function	Potential output	Validation or governance need	Failure risk
Use-case and lifecycle-context definition	Establish scope, stakes, and system setting	Purpose, domain, decisions, users, consequences, lifecycle stage	Prevents context-free debt classification	Context statement	Domain and governance review	Mis-scoped analysis
Stakeholder and decision-context mapping	Identify actors, affected groups, authority, and information needs	Stakeholder map, workflow, decision rights, affected-person analysis	Reveals omitted explanation obligations	Stakeholder–decision map	Meaningful participation review	Missing or tokenized stakeholders
Initial explanation-	Define why, when, for	Tasks, questions,	Creates a baseline for	Explanation-requirement	Human-centered	Generic requirements

requirement definition	whom, and how explanation is needed	expertise, stakes, access needs	identifying deferral	record	evaluation		
Model and data opacity review	Identify inaccessible mechanisms, evidence, provenance, and limitations	Architecture, behavior, data records, tests, transformations	Locates technical and data-opacity debt	Opacity profile	Technical and domain evaluation	Overstated fidelity or data adequacy	
Documentation-gap review	Identify missing, unusable, or stale records	Model, data, system, decision, and explanation documents	Locates documentation debt	Documentation-gap register	Completeness, usability, and currency review	Documentation quantity mistaken for adequacy	
Traceability-gap review	Test links among claims, artifacts, versions, decisions, and actors	Identifiers, lineage, logs, role records	Locates traceability debt	Traceability map	Independent reconstruction	Nominal links without substantive evidence	
Interface-alignment review	Assess fit among explanation, task, user, and context	Interface, workflow, content, accessibility, expertise	Locates interface debt	Alignment findings	Usability and comprehension testing	Attractive but misleading presentation	
Stakeholder-inclusion review	Assess whose needs shaped requirements and evaluation	Participation records and evidence of influence	Locates stakeholder debt	Inclusion-gap analysis	Power, representation, and outcome review	Participation theatre	
Contestability and reviewability assessment	Determine whether decisions can be challenged and examined	Reasons, evidence, records, access, review procedures	Connects debt to governance consequences	Reviewability profile	Institutional and domain assessment	Formal recourse without usable evidence	
Accountability and responsibility mapping	Link lifecycle obligations to actors with authority	Roles, resources, escalation paths, decision rights	Identifies accountability debt	Responsibility map	Organizational-governance review	Blame allocation without effective control	
Governance-risk assessment	Structure pathways from deficits	Debt findings, stakes, controls, uncertainty	Synthesizes lifecycle risk	Qualitative governance-risk narrative	Domain-specific evaluation	False precision or causal overstatement	

	to potential consequences					
Debt register, prioritization, and remediation where supported	Record deficits and organize bounded corrective work	Verified findings, dependencies, urgency, feasibility, owners	Supports transparent attention and correction	Context-specific register and action rationale	Organizational judgment and effectiveness criteria	Unvalidated scoring or activity without improvement
Monitoring, incident learning, and model-update review	Detect change and return lessons to design	Usage evidence, complaints, incidents, drift, version changes	Prevents recurring or stale debt	Monitoring and change-impact findings	Privacy, validity, and release governance	Narrow indicators miss human or institutional failures
Integrated boundary review	Preserve the limits of the conceptual model	Technical, human, organizational, institutional, and residual-risk evidence	Prevents certification or readiness claims	Explicit boundary statement	Independent assessment outside the model	Conceptual findings used as approval

Figure 1 presents the proposed explainability debt model, showing how deferred explanation design, opacity, documentation gaps, traceability deficits, interface misalignment, stakeholder exclusion, and unresolved governance risks accumulate across the AI lifecycle.

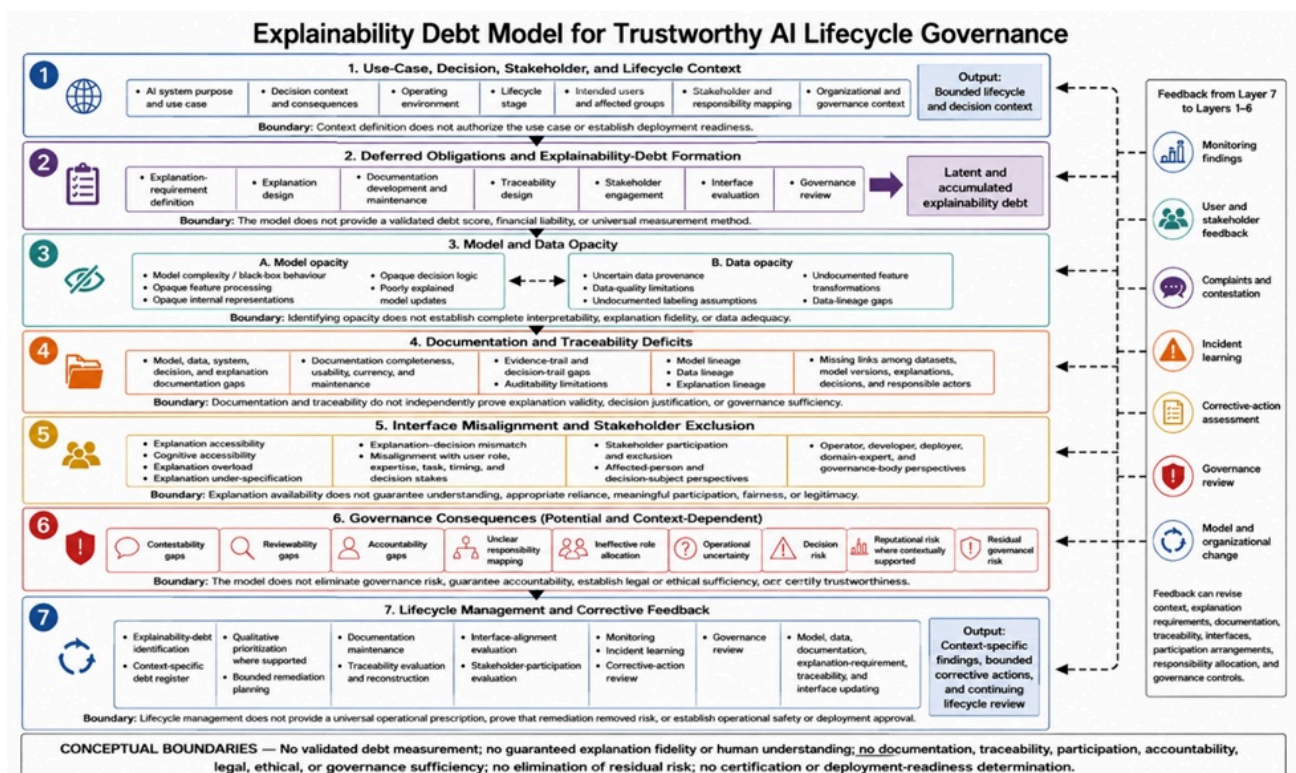


Figure 1. Explainability Debt Model for Trustworthy AI Lifecycle Governance

Figure caption: **Figure 1.** Explainability Debt Model for Trustworthy AI Lifecycle Governance. The model shows how explainability debt can accumulate across the AI lifecycle through deferred explanation design, model opacity, data opacity, documentation gaps, traceability deficits, interface misalignment, stakeholder exclusion, contestability gaps, accountability gaps, unresolved governance risk, monitoring gaps, incident learning gaps, and weak correction feedback. The figure is conceptual and does not represent a validated debt measurement tool, certified governance model, legally sufficient compliance process, or deployment-ready AI management framework.

Alt-text description: A conceptual lifecycle model showing explainability debt accumulating through opacity, documentation gaps, traceability deficits, interface misalignment, stakeholder exclusion, governance risk, monitoring gaps, incident learning gaps, and correction feedback loops.

Management and Research Implications

For management research, the first priority is to examine how explainability debt can be identified and qualitatively prioritized without converting an unvalidated construct into a simplistic score. Studies could investigate how organizations recognize deferred obligations, assign ownership, compare dependencies, and justify remediation sequencing under limited resources. Such work should evaluate whether debt registers improve institutional learning or merely create additional administrative artifacts [27].

Documentation maintenance, traceability reconstruction, interface alignment, and stakeholder participation also require longitudinal evaluation. Research should determine which changes trigger record updates, whether users can reconstruct model and decision histories, how explanation needs vary by expertise and consequence, and whether affected groups influence requirements rather than merely being consulted. These questions concern the effectiveness of lifecycle practices, not the universal adoption of particular templates [16].

Further research should examine accountability operationalization, monitoring, model-update documentation, incident learning, and correction feedback. Relevant questions include whether responsible actors possess authority to intervene, whether explanation failures are detected after deployment, whether incidents change requirements and interfaces, and whether completed corrections reduce the original problem. Organizational capabilities and barriers are likely to shape each of these processes [26].

Explainability-debt measurement remains a longer-term research direction requiring construct definition, domain-specific indicators, reliability testing, stakeholder validation, and evidence that any proposed measure predicts meaningful lifecycle consequences. Until such evidence exists, the model should support conceptual diagnosis and research design rather than legal advice, compliance guidance, procurement decisions, audit instructions, universal documentation or interface prescriptions, technical-debt management instructions, certification, or deployment decisions [11].

Conclusion

This article proposes explainability debt as a conceptual model connecting deferred explanation design with model and data opacity, documentation and traceability deficits, interface misalignment, stakeholder exclusion, governance risk, monitoring gaps, incident learning, and correction feedback across the AI lifecycle. Its central contribution is to show why unexplained AI may create hidden costs that remain difficult to observe until systems must be reviewed, challenged, updated, audited, or governed. Explainability should consequently be treated as a continuing lifecycle responsibility rather than a late-stage add-on, static interface feature, or optional documentation task. The model can help structure analysis of unresolved obligations and feedback failures, but it does not measure debt, prove accountability or trustworthiness, guarantee human understanding, eliminate residual risk, establish legal or ethical sufficiency, certify governance, or determine deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Jul 2025

Revised: 18 Sep 2025

Accepted: 20 Oct 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>.
- 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>.
- Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.
<https://doi.org/10.1007/s12525-020-00441-4>.
- Rios N, Mendonça Neto MG, Spínola RO. A tertiary study on technical debt: types, management strategies, research trends, and base information for practitioners. *Inf Softw Technol*. 2018;102:117-45.
<https://doi.org/10.1016/j.infsof.2018.05.010>.
- Tan J, Feitosa D, Avgeriou P. Does it matter who pays back technical debt? An empirical study of self-fixed TD. *Inf Softw Technol*. 2022;143:106738.

<https://doi.org/10.1016/j.infsof.2021.106738>.

Recupito G, Pecorelli F, Catolino G, Lenarduzzi V, Taibi D, Di Nucci D, et al. Technical debt in AI-enabled systems: on the prevalence, severity, impact, and management strategies for code and architecture. *J Syst Softw*. 2024;216:112151.

<https://doi.org/10.1016/j.jss.2024.112151>.

Moreschini S, Arvanitou EM, Kanidou EP, Nikolaidis N, Su R, Ampatzoglou A, et al. The evolution of technical debt from DevOps to generative AI: a multivocal literature review. *J Syst Softw*. 2026;231:112599.

<https://doi.org/10.1016/j.jss.2025.112599>.

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>.

Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: a review of methods and applications. *Proc IEEE*. 2021;109(3):247-78.

<https://doi.org/10.1109/JPROC.2021.3060483>.

Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89-106.

<https://doi.org/10.1016/j.inffus.2021.05.009>.

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>.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H 3rd, 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.

Königstorfer F, Thalmann S. AI documentation: a path to accountability. *J Responsible Technol*. 2022;11:100043.

<https://doi.org/10.1016/j.jrt.2022.100043>.

Micheli M, Hupont I, Delipetrev B, Soler-Garrido J. The landscape of data and AI documentation approaches in the European policy context. *Ethics Inf Technol*. 2023;25(4):56.

<https://doi.org/10.1007/s10676-023-09725-7>.

Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst*. 2021;11(3-4):1-45.

<https://doi.org/10.1145/3387166>.

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

<https://doi.org/10.1016/j.techfore.2023.122343>.

Ebermann C, Selisky M, Weibelzahl S. Explainable AI: the effect of contradictory decisions and explanations on users' acceptance of AI systems. *Int J Hum Comput Interact*. 2023;39(9):1807-26.

<https://doi.org/10.1080/10447318.2022.2126812>.

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>.

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>.

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>.

Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Sci Eng Ethics*. 2021;27(4):44.
<https://doi.org/10.1007/s11948-021-00319-4>.

Schneider J, Abraham R, Meske C, vom Brocke J. Artificial intelligence governance for businesses. *Inf Syst Manag*. 2023;40(3):229-49.
<https://doi.org/10.1080/10580530.2022.2085825>.

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>.

Padmanabhan Poti S, Stanton CJ. Enabling affordances for AI governance. *J Responsib Technol*. 2024;18:100086.
<https://doi.org/10.1016/j.jrt.2024.100086>.

Porter Z, Zimmermann A, Morgan PDJ, McDermid JA, Lawton T, Habli I. Distinguishing two features of accountability for AI technologies. *Nat Mach Intell*. 2022;4(9):734-6.
<https://doi.org/10.1038/s42256-022-00533-0>.

Subramanian HV, Canfield CI, Shank DB. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. *Artif Intell Med*. 2024;149:102780.
<https://doi.org/10.1016/j.artmed.2024.102780>.