Check for updates

SYSTEMATIC REVIEW

REVISED Factors shaping non-expert trust in AI: A systematic snowballing review of explainable AI design considerations

[version 2; peer review: 1 approved, 1 approved with reservations]

Laura Wright, Iker Legarda, Ganix Lasa, Oscar Escallada

Faculty of Engineering, Design Innovation Centre (DBZ), Mondragon Unibertsitatea, Arrasate, Gipuzkoa, 20500, Spain

v2 **First published:** 20 Apr 2026, **6**:128
https://doi.org/10.12688/openreseurope.22034.1
Latest published: 01 Sep 2026, **6**:128
https://doi.org/10.12688/openreseurope.22034.2

Abstract

Background

Trust in artificial intelligence (AI) systems is critical for ethical and effective adoption, particularly among non-expert users. This review aimed to identify factors influencing trust and practical considerations for designing explainable AI (XAI) systems.

Methods

We systematically reviewed the literature following PRISMA 2020 guidelines and using backward and forward citation searches. The primary search was carried out in Web of Science (WoS), with the final search completed in January 2025. Eligible studies were peer-reviewed publications examining trust in AI among non-expert users and/or design considerations for XAI. Exclusion criteria removed studies focused solely on expert users or unrelated to trust or explainability.

Results

35 studies involving diverse non-expert populations were included. We used an inductive thematic synthesis approach to identify 29 trust-related factors grouped into personal characteristics, sociocultural influences, design and interaction attributes, and system-level properties (e.g., age, emotions, digital literacy, social norms, usability, aesthetics, explanation type, accuracy, transparency, regulatory compliance). Additionally, 120 design considerations were extracted across nine domains, including type and form of explanation; user experience and interaction; trust, transparency and human oversight; robustness, safety and reliability; data, traceability and the AI lifecycle;

Open Peer Review

Approval Status ? ✓

	1	2
version 2 (revision) 01 Sep 2026		✓ view
version 1 20 Apr 2026	? view	? view

1. **Oluwaseun Oladeji Olaniyi**, University of the Cumberlands, Williamsburg, USA

2. **Alexander Hammer**, TU Dresden, Dresden, Germany

Any reports and responses or comments on the article can be found at the end of the article.

regulation and compliance; ethics, values and social responsibility; evaluation methods and metrics; and usage context and organizational factors.

Conclusions

The review synthesizes emerging perspectives on AI trust and explainability and identifies six domains that warrant further research to support the responsible design and deployment of XAI systems. Limitations include heterogeneous evaluation practices and inconsistent operationalization of trust among non-expert users, which complicates comparability across studies and hinder the responsible deployment of XAI systems.

Keywords
User Trust, Trustworthy Artificial Intelligence (TAI), Explainable Artificial Intelligence (XAI), Human–AI Interaction, Snowballing Review

HE This article is included in the Horizon Europe gateway.

Corresponding author: Iker Legarda (ilegarda@mondragon.edu)

Author roles: Wright L: Conceptualization, Data Curation, Formal Analysis, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Legarda I**: Conceptualization, Formal Analysis, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation; **Lasa G**: Funding Acquisition, Project Administration, Supervision, Validation, Writing – Review & Editing; **Escallada O**: Conceptualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101136067(INVERSE).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Wright L *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Wright L, Legarda I, Lasa G and Escallada O. **Factors shaping non-expert trust in AI: A systematic snowballing review of explainable AI design considerations [version 2; peer review: 1 approved, 1 approved with reservations]** Open Research Europe 2026, **6**:128 https://doi.org/10.12688/openreseurope.22034.2

First published: 20 Apr 2026, **6**:128 https://doi.org/10.12688/openreseurope.22034.1

REVISED **Amendments from Version 1**

In this version of the article, we have thoroughly revised the manuscript to address the insightful and constructive feedback provided by the peer reviewers:

- Theoretical Framework (Section 2.2): We added explicit definitions to clarify the distinct roles and conceptual relationships among transparency, interpretability, and explainability within Human-AI Interaction. Furthermore, we incorporated a discussion on the technical requirement of validating XAI predictions and post-hoc explanations (e.g., using perturbation-based methods and fidelity assessments) to ensure explanations accurately reflect true model decision mechanisms rather than ungrounded statistical artifacts.

- Methodological process (Section 4): We expanded the methodological justification for employing a systematic snowballing review procedure to complement single-database searches and mitigate keyword limitations. In addition, we added a dedicated subsection (Section 4.3) detailing the step-wise inductive thematic synthesis procedure.

- Discussion (Section 6): We have completely revised the discussion section. We incorporated Table 7, which links trust factors with design considerations to interpret potential actionable strategies for XAI systems design and evaluation based on the literature. We also strengthened the connection between our proposed "forward-looking perspectives on AI trust and explainability" and the existing literature.

- Structural and Reference updates: Minor edits, updated citations, and structural refinements were implemented across all sections to enhance overall precision, logical flow, and manuscript readability.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

The rapid development of Artificial Intelligence (AI) systems has intensified the scientific debate around their reliability, robustness, and alignment with fundamental principles that enable their responsible integration into sociotechnical environments (de Brito Duarte et al., 2023). In this context, the notion of Trustworthy Artificial Intelligence (TAI) has emerged as a key interdisciplinary concern, bringing together technical, ethical, and cognitive considerations related to how systems behave and how this behaviour is interpreted by users (Hoffman et al., 2023).

Despite broad recognition of its importance, advancing toward truly TAI remains a challenging task. Technical issues, such as the need for more interpretable models, robust risk-management mechanisms, or effective bias-mitigation strategies, coexist with organizational and human-centred difficulties (Brdnik & Šumak, 2024). The trustworthiness of a system critically depends on how end-users understand its capabilities and limitations, the expectations generated by its outputs, and the dynamics that unfold during real-world interactions (Haque, Islam & Mikalef, 2023). These users often interact with AI systems without a detailed understanding of their underlying mechanisms, making them especially vulnerable to misinterpretation, overreliance, or mistrust (Bernardo & Seva, 2023; Habiba et al., 2024; Umm-E-Habiba & Habibullah, 2024).

Within this landscape, advances in Explainable AI (XAI) have become especially relevant, as the explanations accompanying system behaviour can help users understand why a recommendation was produced and evaluate its reliability (Cabour et al., 2021; de Brito Duarte et al., 2023). Despite the increasing research interest in XAI, a significant portion of this research targets technical experts or developers, while the cognitive, emrotional, and contextual factors that shape end-user trust remain insufficiently explored (Haque, Islam & Mikalef, 2023; Myakala, Jonnalagadda & Bura, 2025). As a consequence, practitioners lack actionable guidance for designing XAI systems that effectively support trustworthy interactions for these users.

This article aims to identify the key factors that contribute to gaining and sustaining non-expert end-user trust in AI systems, and to outline design considerations or guidelines for designing XAI systems that support such trust. Through on a review and analysis of recent studies, it seeks to contribute to the unification of existing literature by offering an analytical framework that can guide the development of new AI solutions—or the improvement of existing ones—through a lens of explainability aligned with ethical principles and the real needs of non-expert users.

2. Theoretical framework

This section presents the key concepts required to properly contextualize the subject of study. It defines terms such as trust, TAI, and XAI.

2.1 TAI

Trust is a psychological disposition involving a positive expectation regarding the future behaviour of another agent, even when one is vulnerable to that agent's actions (Rousseau et al., 1998). In the context of AI systems, trust implies the belief that a system will behave in a predictable, safe, and human-aligned manner (Afroogh et al., 2025). However, such trust should not be based solely on superficial impressions or apparent outcomes, but rather on a deep understanding of how and why an AI system makes decisions (Miller, 2019; Liao et al., 2020).

The AI literature has incorporated a taxonomy that distinguishes between different forms of human trust in automated systems. As Jacovi et al. (2021) explain, *intrinsic trust* is grounded in the observable reasoning processes of the AI and the types of explanations it provides. In contrast, *extrinsic trust* arises from external cues, such as the provider's reputation, the perceived quality of the user interface, or the visible outcomes the system delivers. Another relevant distinction is between *justified trust*—where trust in the AI corresponds accurately to its actual ability to fulfil a predictable behavioural contract—and *unjustified trust*, where such alignment is absent and users believe in the system's reliability without sufficient basis. Finally, it is important to distinguish between *cognitive trust*, which is based on a rational analysis of evidence and capabilities, and *emotional trust*, which stems from subjective perceptions formed through interaction, even without technical understanding of the system (Glikson & Woolley, 2020).

The concept of Trustworthy AI (TAI) has been strongly promoted by the European Union, among other international actors, in response to the increasing autonomy and opacity of AI systems. According to the High-Level Expert Group on AI (AI HLEG) of the European Commission, AI is considered trustworthy if it complies with the law, respects fundamental ethical principles, and is safe and reliable throughout its lifecycle (European Commission, 2019). Within the TAI framework, trust is viewed as a contractual transaction (Jacovi et al., 2021), in which a user (trustor) expects the AI system (trustee) to fulfil a set of explicit or implicit expectations (contract).

The need for TAI arises from its growing influence over critical decisions that significantly impact people's lives, such as medical diagnoses, judicial processes, or credit evaluations. Without guarantees of transparency and fairness, there is a risk that these systems will perpetuate injustices or errors, undermining public trust and limiting their beneficial adoption.

Users often trust AI based on its observable outcomes, a form of extrinsic trust. However, this trust can be misleading if the internal decision-making process behind those results is not understood. Intrinsic trust, rooted in knowledge of the system's reasoning and internal functioning, is essential for a comprehensive assessment of its reliability. The absence of intrinsic trust can lead to unjustified trust, where users accept AI decisions without scrutiny, potentially overlooking underlying errors or biases. To prevent this, it is necessary to develop systems that foster both extrinsic trust (through verifiable and consistent outcomes) and intrinsic trust (via explainability, auditability, and human-centred design). Only when both forms of trust converge and are calibrated to the system's actual capabilities can we speak of justified trust and, consequently, of truly trustworthy AI (Vianello et al., 2023).

2.2 XAI

Explainable Artificial Intelligence (XAI) has emerged in response to the increasing complexity and opacity of AI models—particularly those based on deep learning—which often function as "black boxes" that are difficult to interpret. The need to understand and explain the decisions made by these systems has driven the development of XAI, which aims to provide comprehensible interpretations of the internal workings of AI models, thereby promoting transparency and trust in their outputs (Markus et al., 2021). XAI seeks to answer questions such as "Why was this decision made?", "How was this result reached?", and "What factors influenced the outcome?", enabling human users to understand, trust, and effectively manage AI systems. Thus, XAI can be defined as the set of methods and processes that make the outcomes of an AI system understandable to humans without compromising performance (Longo et al., 2024).

For an AI system to achieve meaningful explainability, it fundamentally requires both transparency and interpretability. Transparency refers to a system-level property reflecting the openness, auditability, and overall visibility of the AI pipeline, including its data sources, algorithmic architecture, and operational constraints (Arrieta et al., 2020). In contrast, interpretability represents an intrinsic structural property of a model, denoting the degree to which a human can directly comprehend its internal cause-and-effect decision logic without requiring auxiliary explanatory tools (Miller, 2019). Within this conceptual triad, explainability acts as the active, human-centred communicative bridge that synthesizes system transparency and model interpretability into contextually meaningful representations for non-expert users.

XAI and TAI are intrinsically connected. While TAI focuses on ensuring that AI systems are transparent, robust, and respectful of data privacy, XAI provides the necessary methods and techniques to achieve that transparency and interpretability. In other words, XAI acts as a key enabler for realizing the goals of TAI, allowing users to understand

and justifiably trust the decisions made by AI systems (Chamola et al., 2023). TAI depends on XAI as an operational component, but its full realization also requires the integration of interpretability-oriented design and adherence to normative and ethical principles (Wu et al., 2024).

The reliability of an AI system depends on its accuracy, its transparency, and the user's ability to understand and audit its decisions. XAI enables access to information about how the system processes data, which variables influence its decisions, and how it justifies its predictions (Inam et al., 2021). This strengthens both intrinsic and extrinsic trust, thereby facilitating justified trust—an essential requirement for the adoption of such systems in critical domains such as healthcare, justice, or finance (Cornacchia et al., 2021; Larasati et al., 2023; Miller, 2019).

Crucially, a foundational technical requirement for XAI is the validation of both model predictions and the explanations themselves, as demonstrated in domain-specific technical applications such as multivariate time series classification (Veerappa et al., 2022). Since most XAI techniques rely on post-hoc approximations rather than self-explainable architectures, their explanatory fidelity must be rigorously assessed. Techniques such as perturbation-based methods (e.g., Goettling et al., 2024; Lundberg & Lee, 2017) are essential to verify whether an explanation accurately reflects the model's true decision mechanism. Furthermore, deep neural networks frequently capture non-linear feature combinations that do not naturally map to human reasoning. Therefore, systematic validation is required to ensure that explanations correspond to human-understandable concepts rather than ungrounded statistical artifacts (Hammer et al., 2025).

Several studies have documented the implementation of XAI in sectors such as medicine, where it allows healthcare professionals to interpret diagnostic recommendations; in finance, where it is used to explain credit decisions; and in the autonomous vehicle sector, where it is necessary to justify navigation and behavioural logic (Hernandez et al., 2021; Markus et al., 2021; Samek et al., 2017). These applications demonstrate XAI's potential to enhance the acceptance of intelligent systems and reduce the risk of misinterpretation or misuse.

Nevertheless, despite its expansion across sectors, an integrative framework that systematizes all the factors involved in the reliability of AI systems from the perspective of end users has yet to be developed. This article seeks to identify and thoroughly analyse the factors that scientific literature has highlighted as determinants of end-user trust in AI systems.

3. Research Questions (RQ)

This study has two main objectives: (1) to investigate the key factors that generate and sustain trust among non-expert end-users in AI systems, and (2) to identify the considerations or guidelines to be taken into account when designing new XAI systems or updating existing ones. These two objectives are addressed through the following two research questions respectively:

- RQ1: Which factors influence end-user trust in AI systems?

Motivation: According to Jacovi et al. (2021), trust in AI cannot be understood merely as a technical or algorithmic phenomenon, but rather as a contractual relationship between the user and the system. This relationship depends both on the observable behaviour of the system and the user's understanding of its underlying logic. Other studies have shown that perceived transparency, the clarity of outputs, and the system's level of autonomy are variables that affect the calibration of user trust, potentially leading to either justified or unjustified trust (Glikson & Woolley, 2020). Identifying the specific factors that influence trust—ranging from cognitive to emotional dimensions—is essential for designing AI systems that are both technically efficient and socially acceptable.

- RQ2: What considerations should be taken into account when designing or evaluating an XAI system?

Motivation: The literature highlights that many explanations are tailored to technical audiences and are neither comprehensible nor useful for users without specialized knowledge (Liao et al., 2020). Furthermore, studies such as J. Kim et al. (2024) emphasize the lack of standardized criteria for assessing explanation quality from a user-centred perspective, which hinders the comparability of systems and slows progress toward a unified evaluation practice in XAI. Therefore, it is crucial to identify practical guidelines and design principles that support the development of explainable systems that genuinely respond to users' expectations, capabilities, and contexts.

4. Research methodology

The purpose of this section is to detail the process followed to conduct the systematic literature review. This study adopted a systematic literature review that followed PRISMA guidelines and used a snowballing procedure (i.e. citation and reference searching) due to the emerging, multidisciplinary, and fragmented nature of the research field related to AI system explainability and the factors that influence end-user trust in such systems. This method is particularly suitable

when contributions are dispersed across different disciplines or when theoretical frameworks are not yet well established; it ensures the comprehensiveness, depth, and relevance of the bibliographic corpus (Wnuk & Garrepalli, 2018; Wohlin, 2014).

4.1 Snowballing literature review process

To identify the factors that influence end-user trust in AI systems, the methodological guidelines established by Wohlin (2014) were followed, as illustrated in Figure 1. Wohlin defines the process as an iterative review based on the systematic tracing of citations both backward (looking at papers cited) and forward (looking at papers citing), starting from an initial set of relevant studies (start set).

4.1.1 Selection of the initial set of studies (start set)

Step 1: Database selection and search query. To initiate the search, the Web of Science database was selected due to its usefulness in multidisciplinary fields such as XAI and Human-AI Interaction (HAII), as it integrates publications from computer science, engineering, psychology, social sciences, and technology ethics, among others. This provides a solid and diverse foundation from which to explore the state of the art. Additionally, the search was limited to articles published after 2010 to ensure the relevance and currency of the reviewed studies, given that significant advances in AI—especially in deep learning and autonomous systems—have occurred over the past decade (Azizov et al., 2024).

A set of keywords was identified to locate articles related to the topic of research, resulting in the following search query: "trust" AND "AI" AND "end user". Table 1 presents the rationale for each selected term.

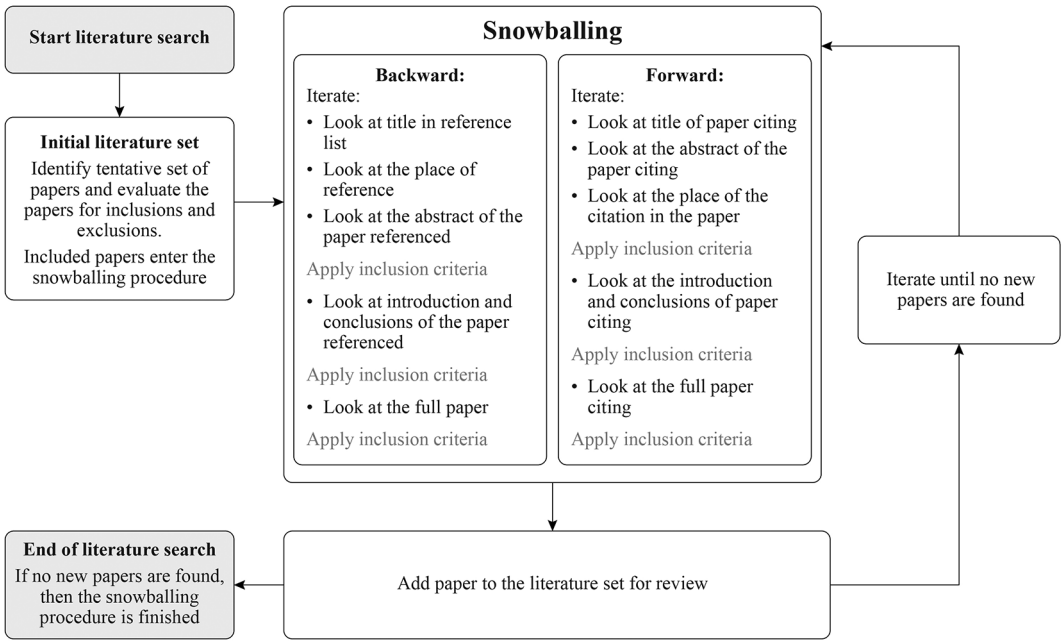

Figure 1. Snowballing literature review procedure. Adapted from Wohlin (2014).

Table 1. Keywords selected for the snowballing literature review.

Keyword	Justification for inclusion
Trust	Central axis of the study, as it examines trust as a critical factor in the adoption and acceptance of AI systems.
AI	Delimits the technological domain of interest, ensuring the search focuses on AI systems rather than other types of technologies.
End user	Focuses the search on studies targeting individuals who interact directly with AI—especially those without technical expertise—thus excluding work aimed solely at developers or technical experts.

Table 2. Inclusion and exclusion criteria used in the snowballing literature review.

Code	Description
IC1	Articles published in English or Spanish.
IC2	Articles published between January 2010 and December 2024.
EC1	Articles that do not address end-user trust or do not include relevant considerations for the design or improvement of XAI systems—for example, studies focused on internal AI mechanisms or developer-oriented explanations.
EC2	Duplicate articles, previously rejected, or already included in the literature review.

Step 2: Selection of the initial set of articles. The literature search was conducted in January 2025. To ensure the currency and relevance of the studies included in this review, the inclusion criteria outlined in Table 2 were applied. Specifically, the selection was restricted to articles written in English or Spanish (IC1) and published between January 2010 and December 2024 (IC2), resulting in a total of 68 articles. After reviewing the titles and abstracts (EC1), 48 articles were excluded. Upon reading the introduction and conclusion of the remaining 20 articles, 14 were retained. Two reviewers independently performed all screening stages, resolving disagreements through discussion, with a third reviewer acting as arbiter when necessary. No automation tools were used.

Finally, the selected articles were read in full and subjected to an in-depth analysis to determine their inclusion in the literature review. Seven articles were excluded, resulting in a final start set of seven articles.

4.1.2 Backward and forward snowballing in iterations

To complement the initial database search and address potential limitations regarding keyword variations and index coverage, a two-iteration backward and forward snowballing approach was systematically applied, starting from the initial set of seven articles. This procedure mitigated the reliance on a single search engine and compensated for potential omissions of relevant studies that employed alternative terminology.

The backward iteration involved examining the reference lists of each article in the initial set, while the forward iteration identified studies that cited those articles (Wohlin, 2014). A three-phase selection procedure was applied in each iteration.

In the first phase, a preliminary screening was conducted based on the title and abstract of each publication, classifying them into three categories: YES, NO, and MAYBE. In the second phase, the introduction and conclusion sections of the articles classified as YES and MAYBE were reviewed in order to more thoroughly assess their theoretical and methodological relevance. Finally, the full texts of the selected articles were read in detail. Only those that met the established inclusion criteria were incorporated as relevant studies for the subsequent iteration of the process. Risk of bias was assessed using the Joanna Briggs Institute (JBI) critical appraisal checklist for systematic reviews (Aromataris et al., 2015).

4.2 Results of the snowballing literature review iterations

Across the two backward snowballing iterations, a total of 1,169 articles were identified and examined. After applying inclusion criteria IC1 and IC2, and conducting a preliminary screening based on titles and abstracts according to exclusion criterion EC1, 1,108 publications were excluded for failing to meet the established requirements. In the second evaluation phase, the introductions and conclusions of the remaining 61 articles were reviewed, leading to the exclusion of 46 articles. Finally, a full reading of the remaining 15 articles resulted in the exclusion of an additional six documents. As a result, nine relevant articles were selected and incorporated into the final corpus through the backward snowballing procedure.

In parallel, forward snowballing retrieved 4,048 articles that cited the initial set. In the first phase, 11 duplicate documents were removed according to exclusion criterion EC2. Subsequently, 3,944 publications were excluded after applying inclusion criteria IC1 and IC2 and conducting an initial screening based on titles and abstracts in line with EC1. In the second phase, the introductions and conclusions of the remaining 93 articles were reviewed, leading to the exclusion of 48 additional papers due to lack of thematic or methodological relevance. A full reading of the 45 remaining papers resulted in the exclusion of another 26. Ultimately, 19 relevant articles were added to the final corpus of this study.

Table 3 presents a summary of the iterations carried out, detailing the number of references and citations reviewed in each round. We stopped after two backward and two forward iterations because the yield of new eligible studies dropped

Table 3. Summary of citations and references screened in each snowballing iteration. Adapted from Wnuk and Garrepalli (2018).

Iteration	FS/BS[1]	Articles rejected (FS)	Articles rejected (BS)	Articles accepted (FS)	Articles accepted (BS)
1	42/418	39	412	3	6
2	4006/751	3994	744	12	7

[1]FS – forward snowballing, BS – backward snowballing.

sharply and no novel concepts emerged. Citation paths increasingly converged on previously included sources, indicating thematic saturation and making further iterations unlikely to substantially change the synthesis. Table 4 shows the final selection of articles included in this snowballing literature review.

4.2.1 Analysis of the literature review results

A total of 35 articles were selected from the two snowballing iterations. Figure 2 illustrates the distribution of articles by year of publication. This distribution reveals a growing trend in scientific output related to the study's topic, with a peak in 2023, during which 11 publications were recorded. In previous years, a moderate progression is observed: 3 articles in 2019, 4 in both 2020 and 2021, followed by a slight increase to 5 in 2022. After the 2023 peak, publication output slightly declined to 7 in 2024 and registered 1 publication in early 2025, retrieved during the forward snowballing iterations. This trend reflects a rising academic interest in explainability, trust, and human-AI interaction, particularly over the last four years.

Table 4. List of articles included in the literature review.

ID	Title	Author(s)	Year
P01	Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective	Bernardo, E; Seva, R	2023
P02	Improving Trustworthiness of AI Solutions: A Qualitative Approach to Support Ethically-Grounded AI Design	Vianello, A; Laine, S; Tuomi, E	2023
P03	An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools	Hernandez, CS; Ayo, S; Panagiotakopoulos, D	2021
P04	Determinants Affecting Consumer Trust in Communication with AI Chatbots: The Moderating Effect of Privacy Concerns	Li, JJ; Wu, LR; Qi, JY; Zhang, YX; Wu, ZY; Hu, SB	2023
P05	Designing for Student Understanding of Learning Analytics Algorithms	Yeh, C; Cowit, N; Howley, I	2023
P06	Improving the User Experience and the Trustworthiness of Financial Services	Cornacchia, G; Narducci, F; Ragone, A	2021
P07	Increasing User Trust in Optimisation through Feedback and Interaction	Liu, J; Marriott, K; Dwyer, T; Tack, G	2022
P08	Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research	Haque, AB; Islam, AKMN; Mikalef, P	2023
P09	Guidelines for Human-AI Interaction	Amershi, S; Weld, D; Vorvoreanu, M; Fourney, A; Nushi, B; Collisson, P; Suh, J; Iqbal, S; Paul N. Bennett, Inkpen, K; Teevan, J; Kikin-Gil, R; Horvitz E	2019
P10	Metrics for Explainable AI: Challenges and Prospects	Hoffman, RR; Mueller, ST; Klein, G; Litman, J	2019
P11	Human-Centered Trustworthy Framework: A Human–Computer Interaction Perspective	Sousa, S; Lamas, D; Cravino, J; Martins, P	2024

Table 4. *Continued*

ID	Title	Author(s)	Year
P12	Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI	Jacovi, A; Marasovic, A; Miller, T; Goldberg, Y	2021
P13	Questioning the AI: Informing Design Practices for Explainable AI User Experiences	Liao, QV; Gruen, D; Miller, S	2020
P14	Human Trust in Artificial Intelligence: Review of Empirical Research	Glikson, E; Woolley, AW	2020
P15	Explainable AI: A Diverse Stakeholder Perspective	Umm-E-Habiba; Habibullah, KM	2024
P16	Towards Explainability as a Functional Requirement: A Vision to Integrate the Legal, End-User, and ML Engineer Perspectives	Umm-E-Habiba; Bogner, J; Wagner, S	2024
P17	Human-centered evaluation of explainable AI applications: a systematic review	Kim, J; Maathuis, H; Sent, D	2024
P18	"Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design	Weitz, K; Schiller, D; Schlagowski, R; Huber, T; André, E	2019
P19	Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users	Larasati, R; de Liddo, A; Motta, E	2023
P20	Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank	Sipos, L; Schäfer, U; Glinka, K; Müller-Birn, C	2023
P21	Four Principles of Explainable Artificial Intelligence	Phillips, P; Han, C; Fontana, P; Broniatowski, D; Przybocki, M	2020
P22	User trust in artificial intelligence: A comprehensive conceptual framework	Yang, R; Wibowo, S	2022
P23	Trust in AI: progress, challenges, and future directions	Afroogh, S; Akbari, A; Malone, E	2025
P24	Trust, distrust, and appropriate reliance in (X) AI: a survey of empirical evaluation of user trust	Visser, R; Peters, T; Scharlau, I; Hammer, B	2023
P25	Explainable AI for Cultural Minds	Kopecka, H; Such, J	2020
P26	Measurable Trust: The Key to Unlocking User Confidence in Black-Box AI	Palazzolo, P; Stahl, B; Webb, H	2024
P27	AI Trust: Can Explainable AI Enhance Warranted Trust?	Duarte, R; Correia, F; Arriaga, P; Paiva, A	2023
P28	A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective	Bach, A; Khan, A; Hallock, H; Beltrao, G; Sousa, S	2022
P29	Invisible Users: Uncovering End-Users' Requirements for Explainable AI via Explanation Forms and Goals	Jin, W; Fan, J; Gromala, D; Pasquier, P; Hamarneh, G	2023
P30	On the Influence of Cognitive Styles on Users' Understanding of Explanations	Riefle, L; Hemmer, P; Benz, C; Vössing, M; Pries, J	2022
P31	Towards an Explanation Space to Align Humans and Explainable-AI Teamwork	Cabour, G; Morales - Forero, A; Ledoux, E; Bassetto, S	2021
P32	Towards Human-centered Explainable AI: User Studies for Model Explanations	Rong; Y; Leeman, T; Nguyen, T; Fiedler, L; Qian, P; Unhelkar, V; Seidel, T; Kasneci, G; Kasneci, E	2022
P33	Exploration of Explainable AI for Trust Development on Human-AI Interaction	Bernardo, EL; Seva, RR	2023
P34	Establishing Appropriate Trust in AI through Transparency and Explainability	Kim, S	2024
P35	Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research	Brdnik S; Šumak B	2024

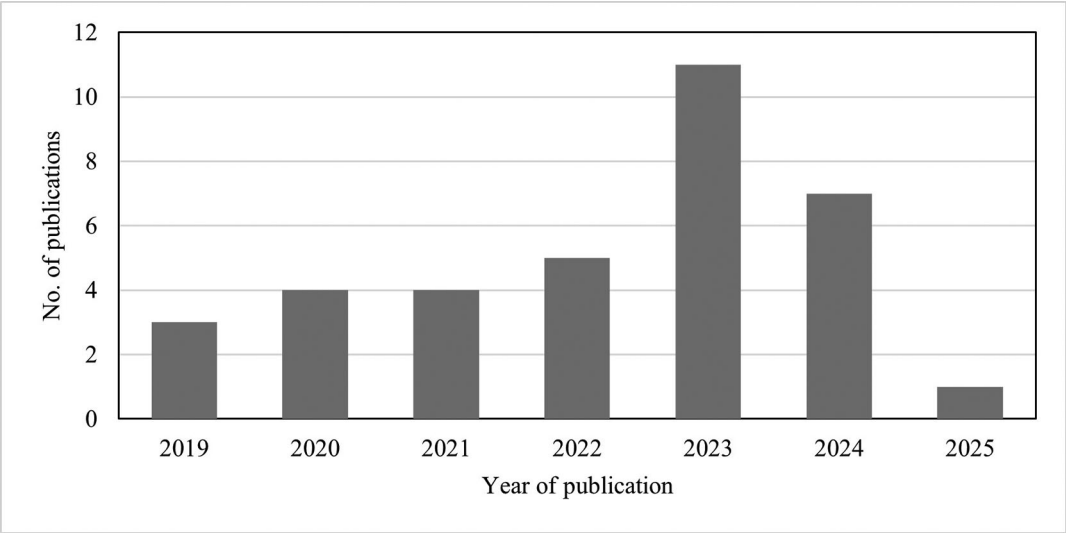

Figure 2. Publications over the years.

Of the 35 articles analysed, 19 were published in scientific journals, making this the primary medium for disseminating research on AI and human trust. Conference proceedings account for 46% of the studies (16 out of 35). Further details on the publication venues included in this literature review are provided in the Extended Data (Wright et al., 2026a).

4.3 Data coding and thematic synthesis

To process and synthesize the qualitative evidence extracted from the 35 selected studies, a step-wise inductive thematic synthesis was conducted. The analytical procedure was structured into two distinct tracks corresponding to the two research questions: the identification and aggregation of trust factors (RQ1), and the extraction and classification of XAI design considerations (RQ2). To ensure coding consistency, transparency, and reproducibility, data extraction and initial coding were performed independently by two researchers. The coders held regular consensus meetings to align code definitions, resolve divergent interpretations through mutual agreement, and minimize individual researcher bias throughout the synthesis process.

4.3.1 Identification and aggregation of trust factors (RQ1)

The identification of factors influencing non-expert trust followed an inductive, bottom-up open coding process. First, trust-related factors explicitly mentioned were recorded. These extracted factors were then iteratively grouped based on conceptual similarity. In a second analytical pass, factors cited in only a single article were re-examined to determine whether they could converge into broader thematic dimensions without losing their core meaning. Convergence was not artificially forced; unique factors were retained as standalone constructs whenever they represented distinct operational or psychological dimensions, such as System Complexity (present in 2 studies) and System Safety Assurance (present in 3 studies). This process yielded a final consolidated set of 29 trust factors.

4.3.2 Extraction, unification, and categorization of design considerations (RQ2)

The synthesis of design guidelines prioritized practical applicability and actionable granularity. An initial set of 345 raw considerations was extracted across the corpus, from which overlapping, synonymous, or phrasing-variant statements were synthesized and unified into a consolidated repertoire of 120 unique design considerations. Within this final set, 54 of the considerations were synthesized from multiple sources, while 66 originated from a single article. A conscious methodological decision was made not to synthesize these guidelines into higher-level abstract statements, thereby preserving their concrete and actionable nature for practitioners. Finally, an inductive thematic grouping exercise was conducted to organize the 120 actionable considerations into nine operational domains, facilitating structured analysis and presentation.

5. Results

This section presents the findings of the thematic synthesis, structured around the two core research questions: the identification of factors shaping non-expert trust (RQ1) and the classification of actionable XAI design considerations (RQ2).

5.1 Factors influencing end-user trust in AI systems (RQ1)

End-user trust in AI systems is a critical component for their responsible adoption, particularly in sensitive or high-impact domains (Afroogh et al., 2025). As previously discussed, Trustworthy AI (TAI) has emerged as a movement strongly promoted by scientific institutions and regulatory bodies as both an ethical and technical priority, aiming to ensure that AI systems are safe, understandable, and aligned with human values (S. Kim, 2024). Within this context, Explainable AI (XAI) has become a key strategy for facilitating the understanding of algorithmic decisions and appropriately calibrating user trust (de Brito Duarte et al., 2023).

Numerous studies have shown that trust does not rely solely on the technical performance of systems, but also on a combination of individual, contextual, ethical, and design-related factors, such as transparency, interpretability, provider reputation, or the type of explanation provided (Bach et al., 2024; Brdnik & Šumak, 2024). Understanding these factors is fundamental to the design of AI systems that foster safer, more comprehensible, and socially acceptable interactions. Table 5 presents the set of 29 trust-related factors identified in the literature, along with a brief description of how each influences end-user trust in AI systems.

Table 5. Factors influencing end-user trust in AI systems.

Ref	Factor	Definition
F01	**Personality**	Individual differences, such as baseline trust tendency or character, affect the predisposition to trust AI systems (Liu et al., 2022; Palazzolo et al., 2024).
F02	**Age**	Age influences how AI trust is perceived and adjusted, as individuals from different generations tend to show varying degrees of familiarity, expectations, and tolerance for technological risk (Liu et al., 2022).
F03	**Gender**	Studies have shown that gender differences can impact both the predisposition to trust and the way trust is calibrated in relation to AI systems (Liu et al., 2022).
F04	**Educational Level**	Educational attainment is positively associated with trust in AI (Yang & Wibowo, 2022).
F05	**Sociocultural Context**	Cultural background influences how system autonomy, acceptable risk, and the legitimacy of delegating decisions to AI are conceptualized. Users from different cultures have distinct trust thresholds shaped by social norms and collective values (Jin et al., 2023; Liu et al., 2022).
F06	**Prior Experience**	A previous positive experience fosters a favourable attitude toward the system, whereas a negative one generates distrust (Cornacchia et al., 2021; Li et al., 2023).
F07	**Digital Literacy**	Users with greater technical understanding perceive less risk, facilitating trust relationships with AI systems (Palazzolo et al., 2024; Yang & Wibowo, 2022).
F08	**Adaptability**	The ability to adjust to individual needs and respond flexibly to change. Tailoring explanations to user context enhances understanding and trust (Cornacchia et al., 2021; Hoffman et al., 2019).
F09	**Anthropomorphism**	Assigning human characteristics to non-human entities. Human-like traits may enhance trust, though excess may lead to the opposite effect (Cornacchia et al., 2021; Li et al., 2023).
F10	**Mental Model**	The internal representation a person has of a system or concept. Understanding these models helps design more effective explanations (Haque et al., 2023).
F11	**Transparency**	The quality of being clear and accessible in the information provided. Systems that clearly explain their processes and decisions foster informed trust by reducing uncertainty and increasing perceived control (Bernardo & Seva, 2023; S. Kim, 2024; Yang & Wibowo, 2022).
F12	**Interpretability**	The quality of being understandable and easy to explain. XAI models should provide clear explanations to help users understand decisions (Amershi et al., 2019; Haque et al., 2023; Liu et al., 2022).
F13	**Interactivity**	The capacity for bidirectional communication between user and system. Allowing users to ask questions or request clarification enhances understanding and trust (Bernardo & Seva, 2023; Hernandez et al., 2021).

Table 5. *Continued*

Ref	Factor	Definition
F14	**Technical Performance**	The system's ability to meet a need or fulfil a purpose. Perceived precision and stability promote trust, especially in critical or high-risk tasks (Afroogh et al., 2025; de Brito Duarte et al., 2023; Visser et al., 2023).
F15	**System Robustness**	The system's ability to operate without failure and resist adverse conditions. Providing consistent, well-founded responses even in complex situations is key to enhancing user trust (Hernandez et al., 2021).
F16	**System Accuracy**	Trust can be assessed by evaluating reliability, accuracy, and potential bias in system performance (Palazzolo et al., 2024).
F17	**Perceived Benevolence**	Users are more inclined to trust systems they perceive as fair and acting in their best interest (Visser et al., 2023; Yang & Wibowo, 2022).
F18	**Regulatory Compliance**	Adherence to current laws and regulations. Ensuring compliance is essential for avoiding legal and ethical risks (Amershi et al., 2019).
F19	**Emotions (anxiety, trust, etc.)**	Affective reactions to stimuli or situations. Accounting for user emotions can enhance explanation acceptance and understanding (Bernardo & Seva, 2023).
F20	**System Design (usability and aesthetics)**	Clear, coherent, user-friendly design enhances perceived credibility and thus trust (Liu et al., 2022).
F21	**System Safety Assurance**	Involves protection from risks or threats. Ensuring data protection and user support boosts trust (Hernandez et al., 2021).
F22	**Application Sector**	Trust levels vary by sector: higher standards are required in healthcare, finance, or justice than in entertainment (Afroogh et al., 2025; Visser et al., 2023).
F23	**System Complexity**	Greater perceived complexity reduces understanding and trust. Algorithmic opacity is a common barrier to sustained trust (Glikson & Woolley, 2020).
F24	**Task Complexity**	When task logic is hard to understand, trust may be limited unless supported by appropriate explanations (Palazzolo et al., 2024; Visser et al., 2023).
F25	**Perceived Use Risk**	The greater the perceived risk (economic, ethical, physical), the higher the user's demand for trust assurances (Liu et al., 2022).
F26	**HAII Type**	The degree of autonomy, control type (manual, supervised, or independent), and perceived agency significantly impact trust development (Glikson & Woolley, 2020).
F27	**System Provider's Reputation**	General opinion based on track record. Backing by reputable organizations reinforces system trust (Hoffman et al., 2019).
F28	**Social Norms and Moral Values (ethics)**	Principles guiding moral behaviour. Users assess trust based on how well the system aligns with their ethical values, justice perceptions, and human rights expectations (Jin et al., 2023; Vianello et al., 2023).
F29	**Feedback**	Information provided about an action to improve performance. Showing intermediate solutions can help build trust, although excessive feedback may induce false confidence (Haque et al., 2023; Vianello et al., 2023).

5.2 Considerations for designing or evaluating an XAI system (RQ2)

The design and evaluation of XAI systems require a differentiated approach from that of conventional AI systems, particularly if the goal is to foster adoption that is safe, comprehensible, and aligned with human values. The concept of TAI specifically emphasizes these qualities, highlighting the need for systems that are not only technically robust but also transparent, auditable, and understandable to end users (S. Kim, 2024). XAI represents a key operational pathway within the TAI framework by offering explanatory mechanisms that enable users to build accurate mental models, calibrate their trust, and exercise informed control over automated decisions (Bach et al., 2024; de Brito Duarte et al., 2023). Ultimately, XAI systems aim to endow AI with qualities that support meaningful, appropriate, and socially responsible human-technology interactions.

Based on this premise, 345 design considerations were identified in the 35 articles reviewed during the systematic synthesis. After a process to eliminate redundancies, the final repertoire was reduced to 120 considerations, grouped into nine categories representing the main dimensions of AI system design and evaluation. Table 6 describes these categories and illustrates several sample considerations, while the full list is available in the Extended Data (Wright et al., 2026a).

Table 6. Categories for considerations for designing or evaluating AI systems.

Category	Definition	Sample considerations
Type and Form of Explanation	What the explanation should display, at what level of granularity, when, and in what format (textual, visual, counterfactual, multimodal, conceptual, etc.). Includes pedagogical and rhetorical principles shaping the explanatory discourse.	*C02 – Consider explanation format as a key design component; C28 – Present intermediate results to allow exploration of solution space*
User Experience and Interaction	Covers usability, personalization, cognitive load, emotional support, adaptability, and feedback mechanisms throughout the usage cycle.	*C03 – Tailor design to perceived system reliability and prior AI experience; C98 – Foster system understanding through mechanisms that improve user–technology interaction*
Trust, Transparency, and Human Oversight	Factors that sustain system credibility, auditability, and the possibility for intervention or correction by experts or end users.	*C85 – Establish accessible accountability mechanisms; C114 – Acknowledge transparency as a key driver of trust among non-expert users*
Robustness, Safety, and Reliability	Measures to ensure stable operation under failures, adversarial attacks, or changing conditions, along with ongoing performance monitoring.	*C44 – Build extrinsic trust using robust and verifiable evaluation schemes; C62 – Design with modularity and adaptability to manage divergent perspectives*
Data, Traceability, and AI Lifecycle	Addresses data quality, process documentation, decision traceability, and long-term maintainability.	*C15 – Implement well-documented learning lifecycles; C73 – Include key content in explanations: condition, treatment, outcomes, processes, and data*
Regulation and Compliance	Adherence to legal frameworks (e.g., GDPR) and audit/accountability requirements imposed by national or supranational directives.	*C63 – Promote a holistic approach integrating technical, legal, and experiential aspects*
Ethics, Values, and Social Responsibility	Covers fairness, bias mitigation, alignment with collective values, cultural inclusion, and corporate social responsibility.	*C06 – Explore trust-building factors through multiple ethical lenses (utilitarian, deontological, virtue ethics); C56 – Consider individual and cultural differences in AI perception*
Evaluation Methods and Metrics	Addresses experimental protocols, including technical validation of post-hoc explanation accuracy and conceptual alignment with human reasoning, objective and subjective metrics, and iterative evaluation strategies with real users.	*C08 – Enable iterative evaluations with real users in real decision contexts; C46 – Assess trust experimentally through causal manipulation*
Usage Context and Organizational Factors	Refers to corporate purpose, anticipated benefits, task–system fit, and sectoral or organizational constraints on deployment.	*C55 – Ensure tasks assigned to the system match its actual capabilities; C112 – Consider subjective task value (Expectancy-Value Theory)*

These considerations address various decision-making levels—from explanation formats to regulatory and organizational frameworks—and reflect a more nuanced understanding of the requirements for genuinely human-centred XAI. Key guidance lines from the analysed studies include:

- **Type and Form of Explanation:** Explanatory design must go beyond merely presenting technical rationales (C05, C33, C35). It is crucial to employ accessible and varied formats (C02, C69), such as multimodal explanations and conceptual representations (C118), tailored to users' cognitive styles (C104, C105) and roles (C24). Explanations should be clear, comprehensible, and grounded in real usage contexts (C22, C49), ideally incorporating examples, counterexamples, and hypothetical scenarios (C50). Interactive explanations are also emphasized (C21), enabling users to explore the system (C28, C76).

- **User Experience and Interaction**: From a UX perspective, both cognitive and affective processes should be activated (C01), while avoiding information overload (C31) and fostering curiosity as a driver of engagement (C38). Iterative human-centred design mechanisms (C25) that incorporate direct user feedback (C119) are also essential. Factors such as calibrated trust (C42, C52), perceived risk (C83), and developer reputation (C19) significantly shape the user–system relationship.

- **Transparency, Human Oversight, and Trust**: Transparency and human oversight form a central axis. Multiple studies emphasize the need for clear accountability mechanisms (C85), explicit trust contracts (C91), and for the system to declare its limitations (C79). Trust should be viewed as dynamic and situated (C40, C72), with evaluation relying on causal manipulation protocols (C46) and validated multidimensional metrics (C87, C88).

- **AI Lifecycle, Regulation, Ethics, and Evaluation**: Concerning the technical and ethical lifecycle of AI, traceability (C13), learning documentation (C15), and ethical principles like fairness, sufficiency, and non-discrimination (C27, C30, C56) are critical. These practices should be supported by comprehensive quantitative and qualitative evaluations (C111) and compliance with legal frameworks such as the GDPR (C108, C99, C63).

- **Usage Context and Organizational Aspects**: Finally, the contextual and organizational nature of explanatory design is emphasized. Explanations should be adapted to the application domain (C101), the system's actual capabilities (C55), and the subjective value users place on tasks (C112). Sociotechnical and ethnographic approaches (C109) can help capture these variables more holistically.

Taken together, these guidelines underscore that explainability should not be reduced to a technical system property. Rather, it must be understood as a relational, situated, and context-sensitive quality. Incorporating these considerations from the early stages of design can lead to AI systems that are more comprehensible, adaptive, and aligned with the expectations, capabilities, and rights of non-expert users.

6. Discussion

This section contextualizes the synthesized findings within broader debates on trust and explainable AI (XAI). To bridge theoretical gaps with practice, the discussion is structured into three interconnected subsections: (1) a critical appraisal of the methodological landscape, domain imbalances, and key empirical limitations; (2) an operational framework mapping 29 trust factors to 120 actionable design and evaluation considerations; and (3) forward-looking perspectives offering our interpretative vision across six emerging research priorities for trustworthy AI.

6.1 Methodological landscape, domain distribution, and critical appraisal of the literature

The analysis of the 35 included studies provides a comprehensive yet heterogeneous overview of how end-user trust and Explainable AI (XAI) are investigated. A critical evaluation of this corpus reveals domain-specific concentration patterns, recurring methodological limitations, and key conceptual gaps in the current state of the art.

6.1.1 Domain Distribution and Scope

The reviewed corpus encompasses both sector-specific empirical evaluations and broader conceptual frameworks. High-stakes application sectors are prominently represented, notably healthcare and clinical decision support (e.g., Larasati et al., 2023; Markus et al., 2021), finance and banking (e.g., Cornacchia et al., 2021), and automotive or autonomous systems (e.g., Hernandez et al., 2021). These are complemented by studies in educational technology (e.g., Yeh et al., 2023) and conversational consumer agents (e.g., Li et al., 2023).

However, two notable structural imbalances emerge from this distribution. First, there is a pronounced over-representation of generic frameworks, as a substantial portion of the literature consists of general Human-Computer Interaction (HCI) frameworks and basic research that lack domain-specific operational grounding (e.g., Amershi et al., 2019; Jacovi et al., 2021; Liao et al., 2020). Second, current research overwhelmingly prioritizes high-stakes environments and high-severity scenarios. Conversely, everyday, low-supervision applications (such as personal productivity recommenders or routine workflow automation) remain understudied, despite the fact that overreliance and subtle trust degradation operate distinctly in lower-risk settings.

6.1.2 Methodological Limitations and Research Gaps

Our synthesis highlights three main methodological limitations characterizing current XAI research targeting non-expert users:

1. *Over-reliance on short-term lab evaluations:* The vast majority of empirical studies assess non-expert trust during brief, single-session laboratory interactions. This cross-sectional approach suffers from inherent novelty effects and artificial task environments, failing to capture how trust dynamically evolves, degrades, or recovers over prolonged, real-world deployment. Crucially, short-term evaluations cannot observe how trust calibration shifts when users encounter unexpected system errors, edge cases, or model updates over time, leaving long-term trust dynamics largely unmeasured.

2. *Inconsistent operationalization and measurement of trust:* Studies frequently conflate subjective user satisfaction, visual appeal, or perceived clarity with true calibrated trust. Empirical research relies heavily on non-standardized, post-hoc questionnaires (such as custom Likert scales) that capture subjective user preference rather than objective, behavioral trust outcomes, including justified compliance, active rejection of incorrect outputs, or user intervention rates. This lack of standardized human-centred metrics hinders cross-study comparability and obscures whether explanations genuinely foster accurate mental models or merely boost user confidence artificially.

3. *Neglect of explanation accuracy and causal validity:* A critical gap in current XAI literature is the tendency to evaluate explanations solely through user-perceived subjective transparency, often ignoring their underlying technical fidelity and real-world ground truth. An explanation generated by post-hoc methods (e.g., LIME, SHAP) may be highly persuasive, intuitive, or visually appealing to a non-expert user while remaining factually inaccurate or causally ungrounded regarding the true domain logic. Because deep learning models frequently rely on statistical correlations or dataset artifacts, an explanation can be algorithmically faithful to a flawed model logic without being causally valid in the real world. This creates a dangerous "illusion of transparency" that induces unwarranted trust and severe overreliance. To foster genuinely justified trust, future research must extend beyond user perceptions by integrating rigorous technical validation through perturbation-based methods and causal attribution tools to verify that explanations convey true cause-and-effect mechanisms before presenting them to end-users.

6.2 Operational framework: Mapping trust factors to considerations for XAI design and evaluation

To bridge the gap between theoretical trust determinants and practical implementation, this section presents an operational framework mapping design considerations directly to the factors influencing non-expert end-user trust. By structuring these relationships as actionable strategies, the framework offers dual utility: providing prescriptive design guidance and establishing tangible criteria for heuristic and empirical evaluation. The structured mapping is presented in Table 7.

In applying this operational lens, individual user characteristics such as personality (F01), age (F02), gender (F03), educational level (F04), and sociocultural context (F05) demand personalized and equitable explanation strategies. Rather than employing uniform interface components, personality profiling enables dynamic adaptation of explanations to match intuitive or rational cognitive styles (C104, C105), while emotional design elements modulate user engagement (C01, C61). Demographic variance in age and education is managed by replacing dense technical metrics with holistic visual metaphors (C61, C100) and embedding pedagogical tooltips (C20, C53) to scaffold domain knowledge (C22, C24). Simultaneously, gender neutrality and fairness criteria (C30, C90) prevent algorithmic bias, while ethnographic grounding (C109) ensures that explanations reflect local socioethical standards (C06, C25, C56, C82, C99).

Beyond baseline demographics, cognitive disposition and user experience shape how explanations are received. Prior experience (F06) and digital literacy (F07) are operationalized through progressive disclosure interfaces (C03) and interactive sandboxes (C113), bridging the gap between actual system reliability and perceived capabilities (C86, C103) while accommodating varying levels of technical familiarity (C51, C96). System adaptability (F08) supports this transition by offering customizable control panels (C32, C62, C102), enabling users to adjust explanation depth dynamically (C36, C57, C84). To foster appropriate social agency without deceptive anthropomorphism (F09), systems carefully integrate conversational traits (C18, C48) and empathetic agent representations (C52, C54, C68). Crucially, these mechanisms align with the user's evolving mental model (F10) by structuring explanations into perception, understanding, and projection layers (C107), while emotional design principles (F19) actively manage user anxiety during system uncertainty (C04, C74, C115).

At the architectural level, the core mechanics of transparency (F11), interpretability (F12), and interactivity (F13) form the functional baseline of explainability. Transparency requires disclosing operational boundaries, inputs, and post-hoc logic (C35, C49, C78, C79, C93), contrasting subjective user confidence with task performance to counter the illusion of transparency (C42, C83). Interpretability mitigates cognitive overload (F23) through concise feature attributions (C02, C31, C100) and hierarchical summaries of complex deep models (C53, C107, C120). Interactivity converts passive viewing into active engagement, allowing end-users to query predictions, override system outputs, and adjust parameters (C21, C28, C34, C119). These elements are orchestrated through iterative human-centred system design (F20), applying rigorous usability testing, task metrics, and standardized evaluation scales (C25, C29, C64, C66, C87, C110, C111, C116).

Technical execution and risk management form another critical pillar in the framework. Technical performance (F14), system robustness (F15), and accuracy (F16) are operationalized via explicit latency displays (C16, C81),

Table 7. Operational mapping of user trust factors and actionable XAI design strategies.

Factor Code	Trust Factor	Mapped Design Considerations	Actionable Strategies for XAI Systems Design and Evaluation
F01	Personality	C01, C03, C38, C57, C61, C72, C104, C105	Design: Implement adaptive UI mechanisms that profile user cognitive styles (C105) and dynamically tailor explanations (e.g., rule-based logic for rational profiles vs. visual/example-based for intuitive ones) (C104). Evaluation: Assess cognitive load and user satisfaction across different personality profiles using A/B testing with varied explanation formats (C01, C87).
F02	Age	C24, C56, C61, C96, C100	Design: Provide multi-tiered explanation detail levels, substituting dense technical metrics with holistic, intuitive visual metaphors for younger or older demographics (C61, C100). Evaluation: Measure comprehension and emotional response across age brackets using usability testing and cognitive walk-throughs (C56, C61).
F03	Gender	C30, C56, C72, C90	Design: Apply strict algorithmic and interface fairness criteria to prevent gender bias, ensuring explanations remain inclusive and demographic-neutral (C30). Evaluation: Audit explanation efficacy and trust levels across gender demographics to verify non-disparity in system utility (C30, C56).
F04	Educational Level	C20, C22, C24, C51, C53, C100	Design: Dynamically calibrate explanation complexity to user educational background, embedding optional pedagogical tooltips to explain foundational domain concepts (C20, C53). Evaluation: Conduct pre/post-intervention knowledge assessments to verify improved domain comprehension across literacy levels (C20, C51).
F05	Sociocultural Context	C06, C25, C30, C56, C82, C99, C109	Design: Incorporate ethnographic studies into the design phase (C109) and localize explanatory terminology to reflect regional cultural paradigms and ethical standards (C56, C99). Evaluation: Conduct cross-cultural field evaluations to measure contextual appropriateness and societal trust acceptance (C06, C109).
F06	Prior Experience	C03, C08, C40, C72, C86, C92, C103	Design: Implement progressive disclosure interfaces that adapt guidance based on system usage history (C03), utilizing visual cues to bridge the gap between actual and perceived system reliability (C86). Evaluation: Perform longitudinal user tracking to evaluate how trust calibration evolves from initial exposure to expert status (C08, C86).
F07	Digital Literacy	C20, C22, C24, C51, C53, C96, C100, C113	Design: Combine direct instructional guides with interactive exploratory sandboxes (C113), prioritizing non-technical, holistic feature representations (C100). Evaluation: Benchmark task completion time and error rates among low digital literacy cohorts using standardized usability scales (C51, C96).
F08	Adaptability	C03, C24, C32, C36, C57, C62, C84, C102	Design: Offer customizable control panels allowing users to adjust explanation depth, modality, and contrast levels (C62, C102). Evaluation: Quantify user preference for manual vs. automated adaptive explanation configurations through log analysis (C32, C102).

Table 7. *Continued*

Factor Code	Trust Factor	Mapped Design Considerations	Actionable Strategies for XAI Systems Design and Evaluation
F09	Anthropomorphism	C04, C17, C18, C48, C52, C54, C68	Design: Carefully integrate conversational traits and human-like interaction patterns without overpromising system capabilities or inducing deceptive trust (C18, C48). Evaluation: Evaluate user perceptions of warmth, competence, and over-reliance using validated social agency questionnaires (C18, C54).
F10	Mental Model	C39, C43, C45, C58, C98, C106, C107	Design: Structure explanations across three distinct layers –perception, understanding, and projection (C107)– to facilitate accurate pattern discovery (C39). Evaluation: Map user mental model fidelity using concept mapping tasks and think-aloud protocols before and after explanation exposure (C39, C107).
F11	Transparency	C04, C05, C17, C33, C35, C42, C49, C58, C59, C60, C73, C77, C78, C79, C83, C89, C91, C93, C114	Design: Disclose model inputs, internal logic, operational boundaries, and potential post-hoc correlation artifacts in real time (C35, C49, C78). Evaluation: Contrast user-perceived subjective transparency with objective task performance to detect and mitigate the "illusion of transparency" and unwarranted overtrust (C78, C83).
F12	Interpretability	C02, C23, C31, C49, C69, C70, C84, C89, C100, C118, C120	Design: Format output representations (e.g., feature attributions, decision trees) to prevent cognitive overload while maintaining information integrity (C02, C31). Evaluation: Test user accuracy and response latency in interpreting system outputs under realistic time constraints (C31, C118).
F13	Interactivity	C21, C28, C32, C34, C38, C98, C103, C119	Design: Build active, bidirectional explanation interfaces enabling users to query predictions, override outcomes, and correct parameters (C21, C34). Evaluation: Track the frequency and efficacy of user overrides and interactive parameter adjustments (C34, C119).
F14	Technical Performance	C12, C16, C45, C46, C55, C75, C80, C81, C97, C115	Design: Make system competence, latency metrics, and service quality indicators explicitly visible within the UI (C16, C81). Evaluation: Correlate system throughput/ response times with subjective user ratings of system capability and professional utility (C16, C81).
F15	System Robustness	C14, C15, C44, C62, C79, C80	Design: Embed explicit boundary alerts and adversarial confidence indicators that signal when inputs approach out-of-distribution limits (C14, C79). Evaluation: Conduct stress-testing and red-teaming scenarios to evaluate user intervention when presented with edge-case system uncertainty (C14, C79).
F16	System Accuracy	C12, C13, C30, C42, C46, C77, C78, C80, C88, C111, C120	Design: Ensure strict end-to-end traceability and causal validity, integrating perturbation-based validation to verify underlying cause-and-effect mechanisms prior to UI presentation (C13, C80). Evaluation: Benchmark explanation fidelity against real-world ground truth and evaluate user fault-detection when exposed to flawed model logic (C13, C80).

Table 7. *Continued*

Factor Code	Trust Factor	Mapped Design Considerations	Actionable Strategies for XAI Systems Design and Evaluation
F17	Perceived Benevolence	C06, C11, C27, C30, C74, C90	Design: Communicate non-maleficence and operational goals clearly (C11), employing empathetic phrasing during high-stakes or sensitive outputs (C74). Evaluation: Administer perceived empathy and organizational trust surveys following critical automated decisions (C11, C74).
F18	Regulatory Compliance	C13, C63, C66, C73, C85, C91, C99, C108	Design: Build automated audit logging and regulatory compliance checks (e.g., GDPR right-to-explanation standards) directly into explanation pipelines (C85, C108). Evaluation: Perform formal legal compliance audits and expert peer reviews on explanation auditability (C85, C108).
F19	Emotions	C01, C04, C61, C74, C115	Design: Apply emotional and reflective design principles to modulate user stress, maintaining supportive visual tone during AI uncertainty (C01, C74). Evaluation: Monitor user physiological stress markers or affective self-assessment grids during system interaction (C01, C74).
F20	System Design	C01, C08, C25, C29, C64, C66, C71, C75, C87, C96, C109, C110, C111, C116	Design: Employ iterative human-centred design cycles, integrating HCI usability principles from early wireframing through deployment (C25, C87). Evaluation: Validate system usability, satisfaction, and cognitive throughput using standardized metrics (e.g., SUS, NASA-TLX) (C87).
F21	System Safety Assurance	C09, C14, C80, C99	Design: Integrate safety-net mechanisms and prominent safety alerts that trigger visual warnings during high-risk automated operations (C09). Evaluation: Simulate safety-critical failures to quantify user emergency response speed and safety protocol adherence (C09, C80).
F22	Application Sector	C07, C65, C101, C108, C109	Design: Domain-adapt explanation taxonomies to conform with sector-specific workflows (e.g., medical diagnostics, financial risk, transport) (C101). Evaluation: Measure domain-expert task efficiency and field-specific trust alignment across targeted deployment sectors (C101, C109).
F23	System Complexity	C23, C31, C53, C100, C107, C110, C120	Design: Deconstruct deep or multi-layered architecture logic into progressive hierarchical summaries (C53, C107). Evaluation: Measure user explanation comprehension speed and subjective complexity ratings across varying model architectures (C53, C107).
F24	Task Complexity	C05, C26, C39, C47, C50, C67, C112	Design: Supply interactive counterfactual explanations, concrete examples, and 'what-if' scenarios to support multi-step decision-making (C26, C50). Evaluation: Benchmark user decision accuracy and problem-solving efficiency in complex vs. simple task conditions (C26, C50).
F25	Perceived Use Risk	C09, C41, C83, C95, C99	Design: Mitigate perceived risk by providing granular user control levers and transparent risk-estimation displays (C83, C95). Evaluation: Conduct risk-perception utility trade-off studies in high-consequence operational environments (C83, C95).

Table 7. *Continued*

Factor Code	Trust Factor	Mapped Design Considerations	Actionable Strategies for XAI Systems Design and Evaluation
F26	HAII Type	C10, C34, C37, C52, C67, C94	Design: Maintain human-in-the-loop control boundaries, providing global overrides tailored to the specific Human-AI Interaction archetype (C10, C37). Evaluation: Evaluate human takeover latency and situational awareness during automated control handoffs (C10, C37).
F27	Provider's Reputation	C19, C82, C91, C117	Design: Establish transparent "trust contracts" and explicitly link interface elements to recognized institutional credentials and organizational standards (C19, C91). Evaluation: Measure pre- and post-interaction institutional trust scores across branded vs. unbranded system variants (C19, C91).
F28	Social Norms & Ethics	C06, C27, C30, C82, C99	Design: Embed socio-ethical guidelines, privacy guarantees, and ethical AI standards throughout the end-to-end interface lifecycle (C27, C99). Evaluation: Assess societal acceptability and ethical compliance through multi-stakeholder feedback frameworks (C27, C99).
F29	Feedback	C08, C28, C36, C76, C103, C119	Design: Integrate continuous human-in-the-loop feedback mechanisms, allowing users to modify parameters and observe real-time model retraining (C76, C119). Evaluation: Quantify model refinement rates and user sense of agency over time through longitudinal feedback logs (C76, C119).

out-of-distribution boundary alerts (C14,C79), and end-to-end traceability (C13,C77,C78). Safety assurance (F21) integrates visual hazard warnings and automated safeguards (C09,C80) to maintain user vigilance. In handling task complexity (F24), interfaces provide interactive counterfactual scenarios (C26,C50) and contextual rationale (C05,C39,C47,C67,C112) to assist multi-step problem-solving. When perceived use risk (F25) is elevated, granular control levers and transparent risk indicators (C41,C83,C95,C99) modulate user trust, while human-AI interaction archetypes (F26) establish explicit control boundaries and override thresholds for automated handoffs (C10,C37,C52,C94).

Beyond technical safeguards, institutional, ethical, and continuous alignment mechanisms embed the system within its operational domain. Perceived benevolence (F17) and provider reputation (F27) are established by clearly articulating operational goals (C11), displaying institutional credentials (C19,C82,C117), and maintaining empathetic communication (C74). Regulatory compliance (F18) and social norms and ethics (F28) embed privacy-by-design, auditable logging, and adherence to legal mandates like GDPR right-to-explanation guidelines (C27,C63,C85,C91,C99,C108). Application sector constraints (F22) dictate domain-adapted explanation taxonomies suited to specific operational workflows (C07,C65,C101). Completing this ecosystem, continuous human-in-the-loop feedback mechanisms (F29) empower users to submit corrective inputs (C08,C36,C76,C103,C119), establishing an iterative retraining loop that maintains calibrated trust over extended real-world deployment.

To illustrate how this framework translates into practical system design, these strategies can be operationalized across distinct high-stakes deployment domains. In clinical decision support interfaces, transparency and accuracy considerations (F11, F16) are operationalized by pairing diagnostic recommendations with visual feature-attribution maps and explicit model uncertainty scores, allowing clinicians to verify underlying medical logic. Similarly, in automated driving displays, robustness and safety requirements (F15, F21) manifest as real-time boundary alerts and intuitive ambient visual cues that warn drivers when edge-case environmental conditions approach automated control limits. Finally, in financial risk assessment dashboards, interactivity and task complexity strategies (F13, F24) are realized through counterfactual 'what-if' simulation panels, enabling analysts to dynamically adjust credit parameters and observe updated risk scores in real time.

Ultimately, synthesizing this operational map makes a significant contribution to bridging the persistent implementation gap between high-level ethical principles and everyday engineering practice. Crucially, this framework is not intended as

a static or definitive artifact, but rather as an extensible starting baseline meant to evolve continuously alongside emerging explainability techniques, domain-specific demands, and novel empirical research. By connecting specific design levers to human and system trust factors, this exercise elevates explainable AI from a post-hoc technical addition to an adaptable, human-centred design discipline. This integrated perspective enables research and development teams to diagnose trust breakdowns in current systems while iteratively refining these actionable strategies to maintain calibrated, justified trust in future applications.

6.3 Forward-looking perspectives on AI trust and explainability

Grounded in the reviewed literature, this section presents our interpretative vision for the next generation of trustworthy AI. We outline six forward-looking perspectives that extend current empirical findings into actionable research priorities and open theoretical questions.

(i) The micro-scale of trust in HAII. Much of the existing research adopts a macro-level approach to trust, focusing on system attributes, user characteristics, or broader contextual elements (Glikson & Woolley, 2020; Visser et al., 2023; Yang & Wibowo, 2022). However, our findings suggest that trust is also shaped through micro-interactions that occur during real-time engagement with the system (Visser et al., 2023). Elements such as response latency, the structure and clarity of information, or the tone of communication can significantly affect how trustworthy the system is perceived to be. For instance, overly fast responses may be interpreted as superficial or insufficiently reasoned, while excessively slow ones may be seen as indicators of inefficiency or incapacity (Bernardo & Seva, 2023). Similarly, subtle aspects like discourse pauses, the granularity of explanations, or the coherence between user input and system output can serve as cues of competence, transparency, or alignment with user expectations (Jacovi et al., 2021; Larasati et al., 2023). This perspective positions trust as a dynamic and emergent phenomenon that is sensitive to the interactional details of use, thus emphasizing the importance of designing and evaluating XAI systems at the micro scale, which in turn will impact the overall user experience.

(ii) The dynamic nature of trust. Trust is neither static nor universal. It evolves in response to both personal and contextual changes. Many of the factors identified in the literature, such as users' educational background, prior experience with AI, or digital literacy, are subject to change over time (Bach et al., 2024; Sousa M et al., 2024). Additionally, broader contextual variables such as regulatory frameworks, cultural norms, or public discourse around AI can shift, altering the landscape in which trust is constructed (Afroogh et al., 2025; Kopecka & Such, 2020; Palazzolo et al., 2024). Even at a day-to-day level, trust may fluctuate depending on the specific task at hand or the user's emotional state or level of fatigue (Bernardo & Seva, 2023; Visser et al., 2023). These dynamics pose challenges for the design of long-term trustworthy AI systems: it is not enough to ensure reliability under current conditions; systems must also be resilient and adaptive in the face of evolving user profiles and usage contexts (Glikson & Woolley, 2020; Sousa M et al., 2024; Visser et al., 2023).

(iii) Context-specific salience of trust factors. Another underexplored dimension is the situational salience of trust-related factors (Vianello et al., 2023). While it is valuable to map the broad range of variables that can influence trust, not all factors carry equal weight in every situation (Bach et al., 2024; Vianello et al., 2023). The same feature (for instance, the tone of interaction) may enhance trust in one context but undermine it in another (Bernardo & Seva, 2023). An informal tone might foster rapport with non-expert users but may be perceived as unprofessional or inappropriate in expert or mission-critical environments (Sanchez Hernandez et al., 2021; Weitz et al., 2019). Therefore, designers must not only consider a wide set of influencing factors but also identify those that are especially significant or even decisive within a given domain, task, or user profile (Liao et al., 2020; Sousa M et al., 2024; Visser et al., 2023). Understanding and prioritizing these high-impact factors is crucial for fostering trust where it matters most.

(iv) Handling errors and recovering trust. A notable blind spot in much of the XAI literature is the implicit assumption that AI systems can or should be error-free. Given the probabilistic nature of many AI models and the complexity of the environments in which they operate, errors are inevitable (Amershi et al., 2019; Sousa S et al., 2024). In this regard, insights from service science, particularly the concept of service recovery, may offer valuable guidance. Service recovery refers to the strategies deployed to regain user trust following a failure (Van Vaerenbergh & Orsingher, 2016). Prior research shows that a well-managed recovery process can enhance user satisfaction to levels higher than before the error occurred (McCollough & Bharadwaj, 1992). For AI systems, such recovery mechanisms might include post-hoc explanations, automated apologies, opportunities for user correction, or escalation to human oversight. Understanding how trust is damaged and how it can be restored is a critical area of inquiry that remains underdeveloped in the current discourse on explainable AI (Jacovi et al., 2021; Visser et al., 2023).

(v) Social and relational dimensions of trust in AI. Another significant gap in the literature is the limited attention to the social and relational dimensions of trust in human-AI interaction. In related fields such as social robotics or human-robot

interaction (HRI), it has been shown that trust is shaped not only by technical competence but also by relational cues (Glikson & Woolley, 2020; Jacovi et al., 2021; Weitz et al., 2019). Concepts such as self-other recognition, altruistic behaviour, and the ability to predict human intentions are crucial to the formation of trust in these domains (Afroogh et al., 2025; Jacovi et al., 2021). As AI systems increasingly assume collaborative and proactive roles, particularly in settings that demand teamwork or shared decision-making, these relational dimensions become ever more relevant (Afroogh et al., 2025; Cabour et al., 2021; Glikson & Woolley, 2020). Treating AI solely as a functional tool risks overlooking how users emotionally and socially engage with these systems, especially in high-stakes or long-term interactions.

(vi) Human-centred evaluation methods for trustworthy AI. Finally, there is a clear need for evaluation frameworks that are grounded in human factors, with a particular focus on trust. While technical metrics and algorithmic transparency are important, they offer an incomplete picture of how trust is experienced by end users (Rong et al., 2022; Visser et al., 2023). Cross-disciplinary insights from human-computer interaction and psychophysiology suggest a range of potentially applicable methods, from heuristic evaluations and usability testing (Amershi et al., 2019; Liao et al., 2020) to biometric techniques such as heart rate variability, electrodermal activity, brain activity, or eye-tracking (Fairclough, 2009; Lu & Sarter, 2019; Visser et al., 2023). These methods can provide more nuanced and less biased insights into how trust is formed, maintained, or lost (Afroogh et al., 2025; S. Kim et al., 2024; Visser et al., 2023). Adapting and validating such approaches for AI interaction contexts could significantly advance our capacity to assess and optimize trust-related aspects of system design (Bach et al., 2024; Sousa M et al., 2024; Visser et al., 2023).

7. Conclusions and Future Directions

This study has presented a systematic literature review conducted using the snowballing method, with the aim of identifying, synthesizing, and connecting two key dimensions in the field of explainable and trustworthy AI: on the one hand, the factors that influence end-user trust in AI systems (RQ1), and on the other, the relevant considerations for the design and evaluation of XAI systems (RQ2).

The findings have made it possible to construct a comprehensive and structured view of trust as a multifactorial, situated, and dynamic phenomenon, one whose consolidation depends both on individual user characteristics and on the technical, communicative, and ethical design of the system. The identification of 29 trust-related factors and 120 design considerations, extracted from a refined corpus of 35 reviewed articles, provides an empirical foundation for advancing toward design frameworks that are more responsive to the diversity of users and contexts.

The use of the snowballing approach has enabled the coherent mapping of a highly relevant and interconnected body of scientific literature. Moreover, the results presented lay the groundwork for future research aimed at expanding, validating, and operationalizing the identified links between trust, explainability, and user experience. This review contributes to the consolidation of a perspective that regards explanatory design as a central practice for ensuring the ethical, comprehensible, and safe adoption of AI systems.

7.1 Theoretical contributions

The findings of this research provide a significant theoretical contribution to the fields of XAI and TAI by coherently articulating the factors that influence user trust with the design and evaluation considerations necessary for explainable systems. Crucially, this is operationalized through a novel mapping framework (Table 7) that directly pairs trust determinants with actionable design strategies. This articulation moves beyond the fragmented approaches that have dominated the literature in recent years, where normative frameworks remain disconnected, by proposing an integrative vision that conceptualizes trust as a multidimensional construct situated at the intersection of individual, contextual, and design-related variables.

In this regard, trust is not conceived as an inherent property of the system, nor as a mere outcome of its technical performance, but rather as a dynamic and relational process that emerges from the interaction between the user, the system, and the context of use. This perspective aligns with recent approaches in HCI and with cognitive and ethical theories that emphasize the importance of factors such as prior experience, risk perception, transparency, or cultural alignment. Integrating these factors into a systematic framework advances the development of a situated theory of trust in AI systems, where explanatory design plays a central role as a social and affective mediator.

The combination of trust factors and design considerations offers a conceptual bridge between the abstract principles of TAI (such as fairness, safety, explainability, or human oversight) and their translation into concrete, specific, and adaptable design practices for XAI systems. This approach helps address a recognized gap in the literature: the challenge of operationalizing ethical values in a practical and assessable way in real-world usage contexts**.**

This research contributes to consolidating a richer, more critical, and user-oriented vision of XAI, in which design becomes a structural component in the development of AI systems that aim to be socially legitimate, culturally sensitive, and humanly understandable.

7.2 Practical contributions

In terms of practical implications, this study proposes a structured mapping between trust factors and design considerations that serves as a set of guiding criteria for decision-making during the early design phases, particularly in defining user profiles, conducting task analysis, and modelling the context of use, as well as in the review and evaluation of existing systems. This operational bridge demonstrates how abstract trust goals translate into practical UI/UX mechanisms across high-stakes domains, such as paired feature attributions in clinical decision support, real-world boundary alerts in automated driving, or interactive 'what-if' counterfactual simulations in financial risk assessment.

From an interaction design perspective, the results highlight the importance of generating explanatory experiences that are adaptive, interactive, and emotionally coherent. This requires moving beyond the paradigm of "static explanation" toward activity-centred models that enable users to explore, compare, intervene in, and, if necessary, correct the system's decisions.

Regarding evaluation, the study proposes tools to operationalize trust as an observable variable, incorporating both objective indicators (such as reliability, consistency, and traceability) and validated subjective measures (such as comprehension, satisfaction, perceived risk, and perceived usefulness).

Moreover, in sectors where AI is implemented in critical tasks or with vulnerable populations, this research offers concrete principles to ensure the protection of rights, equity, and the safety of end users. The integration of ethical principles throughout the design lifecycle, transparent communication of system limitations, and explicit compliance with regulatory frameworks become practically indispensable for reducing the risk of unfair, biased, or misunderstood decisions.

The study underscores that trust is not an emergent property of technical models, but rather a design responsibility. Trust must be designed, mediated, and evaluated based on the quality of interactions, the clarity of communication, the ethical sensitivity of the system, and its ability to align with users' values, expectations, and vulnerabilities.

7.3 Limitations

This research presents a series of limitations that must be acknowledged, both to properly frame the scope of the conclusions and to guide future lines of inquiry. It is important to note that the domains of XAI and TAI constitute a dynamic, expansive, and still conceptually emerging field, shaped by the convergence of approaches from multiple disciplines, including computer science, psychology, philosophy, requirements engineering, and human-centred design (HCD). This plurality of perspectives results in fragmentation, which complicates the development of unified and comprehensive guidelines, particularly regarding the design of AI systems adaptable to diverse users, contexts, and capabilities. While this study proposes an integrative approach linking trust factors and design considerations, a fully operational theory of explanatory adaptability has yet to be developed. Such a theory would need to systematically address variables such as the cognitive level, sociocultural background, emotional load, and users' technological experience.

Secondly, prioritizing general user terms (e.g., "end-user", "non-expert") over role-specific titles (e.g., "clinician", "driver", "operator") may have omitted empirical studies referencing users solely by job designations who, despite domain expertise, remain AI non-experts. Consequently, the synthesis highlights cross-domain design principles rather than sector-specific operational requirements.

Finally, specific methodological trade-offs in the literature search strategy must be explicitly recognized. The primary retrieval phase relied on a single database and a targeted set of keywords, which may have excluded relevant studies that employ alternative terminology or are indexed in other platforms. To mitigate these initial restrictions and compensate for potential search omissions, a systematic two-iteration backward and forward snowballing strategy was applied. However, this citation-dependent approach carries its own inherent limitations. The analysed corpus may not fully capture the most recent or emerging literature, especially that published in interdisciplinary conferences or in non-traditionally indexed formats. Moreover, the selection process is influenced by the initial starting points and the inclusion criteria applied to related studies, which may introduce a degree of thematic or methodological bias. The snowballing strategy inherently depends on citation networks, which may omit unpublished studies or research with unreported or negative findings, potentially limiting the completeness of the evidence base. Although measures were taken to ensure the quality and

representativeness of the document set analysed, it cannot be claimed that the set of extracted considerations or influencing factors is exhaustive or definitive.

7.4 Future research lines

Building on the limitations identified and the forward-looking perspectives outlined in the discussion section, several avenues for future research emerge that could help consolidate and expand the findings presented here.

A key priority is the advancement toward the formulation of comprehensive, operational, and adaptable design guidelines that can be applied across various sectors. These guidelines should link individual factors to specific considerations and enable the integration of the five levels of complexity involved in explanatory design: cognitive, affective, ethical, technical, and situational. To achieve this, it will be necessary to delve deeper into how each of the identified trust factors can be addressed with new, more specific considerations beyond those extracted in this review. This development could be enriched through qualitative interviews with experts in human-AI interaction, usability, algorithmic ethics, or technology regulation, allowing for the collection of expert knowledge not published or not yet structured in the academic literature.

Additionally, this work would benefit from being complemented by participatory empirical studies that incorporate the perspectives of end users with varying levels of experience, capabilities, and cultural frameworks. Such studies would allow for the validation of the proposed considerations under real-world conditions, help identify potential barriers to implementation, and support the adaptation of strategies to specific domains. In parallel, a broader approach is recommended, including longitudinal validations and studies that examine how trust evolves over time and in relation to different forms of representation, transparency, and perceived agency. This would contribute to building a more situated, dynamic, and context-sensitive understanding of trust in AI systems, an essential element for a socially robust and legitimate XAI aligned with the values of Trustworthy AI.

Furthermore, empirical research must investigate how trust is constructed through real-time micro-interactions (e.g., system response latency, pauses, and communicative tone) and explore how trust factors shift in situational salience depending on task criticality. A particularly critical and underdeveloped line of inquiry involves evaluating system error recovery; future studies should adapt service recovery principles from service science to explore post-hoc explanations, automated apologies, and human-in-the-loop overrides as mechanisms to restore damaged trust. As AI systems assume increasingly collaborative and proactive roles, future work must also address the relational and social dimensions of trust, drawing on insights from human-robot interaction.

Moreover, to move beyond subjective user preference questionnaires and address the illusion of transparency, future research should develop human-centred evaluation frameworks incorporating objective, neurophysiological, and biometric measurement tools. Integrating eye-tracking, heart rate variability, and electrodermal activity can provide unobtrusive, real-time insights into cognitive load and affective state. In addition, future studies should integrate systematic validation techniques, such as perturbation-based methods and weight-randomization sanity checks, to verify whether explanations of complex model features genuinely align with human-understandable concepts rather than merely triggering user expectation bias. Combining such conceptual fidelity checks with objective biometric tools will ultimately establish more rigorous benchmarks for calibrated trust.

Finally, a strategic line of research involves the development of methodological tools derived from this framework, such as decision matrices, self-assessment instruments for design teams, or support systems for selecting explanatory formats based on user profiles. These tools could facilitate the incorporation of trust and explainability principles from the early stages of technological development, reducing the risk of solutions focused solely on technical or functional criteria.

Ethics and consent statement

Ethical approval and consent were not required for this study, as it is a systematic literature review based exclusively on previously published studies.

Data availability statement

The underlying and extended data has been deposited in the eBiltegia repository of Mondragon Unibertsitatea as “Systematic Literature Review Data on End-User Trust and Explainability in AI Systems”. DOI: https://doi.org/10.48764/2qk2-pv48 (Wright et al., 2026a).

This dataset contains the following data:

- Data file 1: "Research Data_SLR End-user trust and explainability in AI systems.xlsx" (Underlying and extended data of the Systematic Literature Review process)
- Data file 2: "Research Data_Factors influencing end-user trust in AI systems and design considerations.xlsx" (Underlying and extended data of the Data analysis and results)
- Data file 3: "Publication Venues of the Selected Articles" (Extended data of the Systematic Literature Review results)
- Description of the dataset: "Readme Research Data_SLR End user trust and explainability in AI systems.txt"

Furthermore, the PRISMA checklist and flow diagram used to ensure methodological rigor in the present article have been deposited in the eBiltegia repository of Mondragon Unibertsitatea as "Non-Expert Trust in AI: PRISMA checklist and flow diagram". DOI: https://doi.org/10.48764/C2VB-P608 (Wright et al., 2026b).

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

References

Afroogh S, Akbari A, Malone E, *et al.*: **Trust in AI: progress, challenges, and future directions.** *Humanit. Soc. Sci. Commun.* 2025; **12**(1): 67.
Publisher Full Text

Amershi S, Weld D, Vorvoreanu M, *et al.*: **Guidelines for Human-AI Interaction.** *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 2019; pp. 1–13.
Publisher Full Text

Aromataris E, Fernandez R, Godfrey CM, *et al.*: **Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach.** *Int. J. Evid. Based Healthc.* 2015; **13**(3): 132–140.
PubMed Abstract | **Publisher Full Text**

Arrieta AB, Díaz-Rodríguez N, Del Ser J, *et al.*: **Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.** *Information fusion.* 2020; **58**: 82–115.
Publisher Full Text

Azizov D, Manzoor MA, Bojkovic V, *et al.*: **A Decade of Deep Learning: A Survey on The Magnificent Seven.** 2024.
Reference Source

Bach TA, Khan A, Hallock H, *et al.*: **A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective.** *Int. J. Hum. Comput. Interact.* 2024; **40**(5): 1251–1266.
Publisher Full Text

Bernardo E, Seva R: **Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective.** *Informatics.* 2023; **10**(1): 32.
Publisher Full Text

Brdnik S, Šumak B: **Current Trends, Challenges and Techniques in XAI Field: A Tertiary Study of XAI Research.** *2024 47th ICT and Electronics Convention, MIPRO 2024 - Proceedings.* 2024; pp. 2032–2038.
Publisher Full Text

Cabour G, Morales A, Ledoux É, *et al.*: **Towards an Explanation Space to Align Humans and Explainable-AI Teamwork.** 2021.
Reference Source

Chamola V, Hassija V, Sulthana AR, *et al.*: **A Review of Trustworthy and Explainable Artificial Intelligence (XAI).** *IEEE Access.* 2023; **11**: 78994–79015.
Publisher Full Text

Cornacchia G, Narducci F, Ragone A: **Improving the User Experience and the Trustworthiness of Financial Services.** 2021; 264–269.
Publisher Full Text

de Brito Duarte R, Correia F, Arriaga P, *et al.*: **AI Trust: Can Explainable AI Enhance Warranted Trust?** *Hum. Behav. Emerg. Technol.* 2023; **2023**: 1–12.
Publisher Full Text

European Commission: *A definition of AI: Main capabilities and scientific disciplines.* The European Commission's High-Level Expert Group on Artificial Intelligence; 2019.

Fairclough SH: **Fundamentals of physiological computing.** *Interacting with computers.* 2009; **21**(1-2): 133–145.
Publisher Full Text

Glikson E, Woolley AW: **Human Trust in Artificial Intelligence: Review of Empirical Research.** *Acad. Manag. Ann.* 2020; **14**(2): 627–660.
Publisher Full Text

Goettling M, Hammer A, Malberg H, *et al.*: **xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features.** *Scientific Reports.* 2024; **14**(1): 13122.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Habiba U-E, Bogner J, Wagner S: **Towards Explainability as a Functional Requirement: A Vision to Integrate the Legal, End-User, and ML Engineer Perspectives.** *Proceedings of the 2nd International Workshop on Responsible AI Engineering.* 2024; pp.16–19.
Publisher Full Text

Hammer A, Goettling M, Malberg H, *et al.*: **Fusion of automatically learned rhythm and morphology features matches diagnostic criteria and enhances AI explainability.** *npj Artificial Intelligence.* 2025; **1**(1).
Publisher Full Text

Haque AB, 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. Chang.* 2023; **186**: 122120.
Publisher Full Text

Hernandez CS, Ayo S, Panagiotakopoulos D: **An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools.** *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC).* 2021; 1–10.
Publisher Full Text

Hoffman RR, Mueller ST, Klein G, *et al.*: **Metrics for Explainable AI: Challenges and Prospects.** *Comput. Therm. Sci.* 2019.

Hoffman RR, Mueller ST, Klein G, *et al.*: **Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance.** *Front. Comp. Sci.* 2023; **5**.
Publisher Full Text

Inam R, Terra A, Mujumdar A, *et al.*: **Explainable AI – how humans can trust AI.** 2021.
Reference Source

Jacovi A, Marasović A, Miller T, Goldberg Y: **Formalizing Trust in Artificial Intelligence.** *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 2021; pp. 624–635.
Publisher Full Text

Jin W, Fan J, Gromala D, *et al.*: **Invisible Users: Uncovering End-Users.** *Requirements for Explainable AI via Explanation Forms and Goals.* 2023.
Publisher Full Text

Kim J, Maathuis H, Sent D: **Human-centered evaluation of explainable AI applications: a systematic review.** *Front. Artif. Intell.* 2024; **7**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kim S: **Establishing Appropriate Trust in AI through Transparency and Explainability.** *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* 2024; pp. 1–6.
Publisher Full Text

Kopecka K, Such JM: **Explainable AI for Cultural Minds.** *Paper presented at* Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction. 2020.

Larasati R, De Liddo A, Motta E: **Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users.** *ACM Trans. Interact. Intell. Syst.* 2023; **13**(4): 1–39.
Publisher Full Text

Li J, Wu L, Qi J, *et al.*: **Determinants Affecting Consumer Trust in Communication With AI Chatbots.** *J. Organ. End User Comput.* 2023; **35**(1): 1–24.
Publisher Full Text

Liao QV, Gruen D, Miller S: **Questioning the AI: Informing Design Practices for Explainable AI User Experiences.** *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 2020; pp. 1–15.
Publisher Full Text

Liu J, Marriott K, Dwyer T, *et al.*: **Increasing User Trust in Optimisation through Feedback and Interaction.** *ACM Trans. Comput. Hum. Interact.* 2022; **29**(5): 1–34.
Publisher Full Text

Longo L, Brcic M, Cabitza F, *et al.*: **Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions.** *Inf. Fusion.* 2024; **106**(February): 102301.
Publisher Full Text

Lu A, Sarter N: **Eye-tracking metrics for measuring trust in automation.** *Human Factors.* 2019; **61**(7): 1167–1182.
Publisher Full Text

Lundberg S, Acirc M, Lee S-I: **A unified approach to interpreting model predictions.** *Adv. Neural Inf. Process. Syst.* 2017; **30**.
Publisher Full Text

Markus AF, Kors JA, Rijnbeek PR: **The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies.** *J. Biomed. Inform.* 2021; **113**: 103655.
PubMed Abstract | **Publisher Full Text**

McCollough MA, Bharadwaj SG: **The recovery paradox: an examination of consumer satisfaction in relation to disconfirmation, service quality, and attribution based theories** *Marketing Theory and Applications.* 1992; **119**(3).

Miller T: **Explanation in artificial intelligence: Insights from the social sciences.** *Artif. Intell.* 2019; **267**: 1–38.
Publisher Full Text

Myakala PK, Jonnalagadda AK, Bura C: **The Human Factor in Explainable AI Frameworks for User Trust and Cognitive Alignment.** *Int. Adv. Res. J. Sci. Eng. Technol.* 2025; **12**(01): 80–93.
Publisher Full Text

Palazzolo P, Stahl B, Webb H: **Measurable Trust: The Key to Unlocking User Confidence in Black-Box AI.** *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems.* 2024; pp. 1–7.
Publisher Full Text

Rong Y, Leemann T, Nguyen TT, *et al.*: **Towards Human-centered Explainable AI: User Studies for Model Explanations.** 2022.
Publisher Full Text

Rousseau DM, Sitkin SB, Burt RS, *et al.*: **Not So Different After All: A Cross-Discipline View of Trust.** *Acad. Manag. Rev.* 1998; **23**(3): 393–404.
Publisher Full Text

Samek W, Wiegand T, Müller K-R: **Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.** 2017.
Publisher Full Text

Sousa S, Lamas D, Cravino J, *et al.*: **Human-centered trustworthy framework: a human–computer interaction perspective.** *Computer.* 2020; **57**(3): 46–58.

Sousa M, Aresta M, Pedro L, *et al.*: **Human-centered XAI design: Operationalizing dynamic trust factors across multi-context AI implementations.** *Int J Hum Comput Stud.* 2024; **181**: 103150.
Publisher Full Text

Umm-E-Habiba, Habibullah KM: **Explainable AI: A Diverse Stakeholder Perspective.** *2024 IEEE 32nd International Requirements Engineering Conference (RE).* 2024; pp. 494–495.
Publisher Full Text

Van Vaerenbergh Y, Orsingher C: **Service recovery: An integrative framework and research agenda.** *Acad. Manag. Perspect.* 2016; **30**(3): 328–346.
Reference Source

Veerappa M, Anneken M, Burkart N, *et al.*: **Validation of XAI explanations for multivariate time series classification in the maritime domain.** *Journal of Computational Science.* 2022; **58**: 101539.
Publisher Full Text

Vianello A, Laine S, Tuomi E: **Improving Trustworthiness of AI Solutions: A Qualitative Approach to Support Ethically-Grounded AI Design.** *Int. J. Hum.-Comput. Interact.* 2023; **39**(7): 1405–1422.
Publisher Full Text

Visser R, Peters T, Scharlau I, *et al.*: **Trust, distrust, and appropriate reliance in (X)AI: a survey of empirical evaluation of user trust.** *Cognitive Interaction Technology.* 2023.
Publisher Full Text

Weitz K, Schiller D, Schlagowski R, *et al.*: **"Do you trust the AI?" The effect of explanation modalities on user trust in AI system recommendations.** *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19).* 2019: 79–86.
Publisher Full Text

Wnuk K, Garrepalli T: **Knowledge management in software testing: A systematic snowball literature review.** *E-Informatica Softw. Eng. J.* 2018; **12**(1): 51–78.
Publisher Full Text

Wohlin C: **Guidelines for snowballing in systematic literature studies and a replication in software engineering.** *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering.* 2014; pp. 1–10.
Publisher Full Text

Wright L, Legarda I, Lasa G, *et al.*: Systematic Literature Review Data on End-User Trust and Explainability in AI Systems (Version 1). [Dataset]. EBiltegia Repository, Mondragon University; 2026a.
Publisher Full Text

Wright L, Legarda I, Lasa G, *et al.*: Non-Expert Trust in AI: PRISMA checklist and flow diagram (Version 1). [Dataset]. EBiltegia Repository, Mondragon University; 2026b.
Publisher Full Text

Wu C, Buyya R, Li YF, *et al.*: Mindset = Explainable AI. Dataset. Interpretable AI; 2024.
Reference Source

Yang R, Wibowo S: **User trust in artificial intelligence: A comprehensive conceptual framework.** *Electron. Mark.* 2022; **32**(4): 2053–2077.
Publisher Full Text

Yeh YC, Wu JY, Chen SY: **Influence of XAI-based pedagogical agents on students' trust and learning engagement in AI-supported educational technology.** *Computers & Education.* 2023; **194**: 104689.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓

Version 2

Reviewer Report 10 September 2026

https://doi.org/10.21956/openreseurope.26520.r80116

© **2026 Hammer A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ **Alexander Hammer**
TU Dresden, Dresden, Germany

Thank you for considering me to review the revised manuscript. The authors have incorporated all my suggestions to my complete satisfaction and, in some places, even made sensible additions. Therefore, I have no further comments and recommend the acceptance of the manuscript without further revision.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 13 June 2026

https://doi.org/10.21956/openreseurope.23842.r74029

© **2026 Hammer A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Alexander Hammer**
TU Dresden, Dresden, Germany

The submitted manuscript presents a snowballing review of the current research landscape regarding factors that influence non-expert trust in artificial intelligence (AI) systems and the application of explainable AI (XAI) in this context. The authors employ a PRISMA-based systematic

literature search, followed by a snowballing methodology, to identify and analyze relevant articles. The review synthesizes both user trust factors and design considerations for XAI, offering a comprehensive mapping of the field. The topic is highly relevant, and the manuscript is, overall, well-structured and clearly written.

Major Comments

General

[1] The terms *transparency*, *explainability*, and *interpretability* have distinct meanings within XAI literature, yet the manuscript sometimes uses them interchangeably and at other times does not. For clarity and academic rigor, these terms should be explicitly defined at the outset and used consistently throughout the paper.

[2] The technical validation of AI systems is a critical foundation for trustworthiness, yet this aspect is underrepresented in the manuscript. This includes the **AI model validation** (The accuracy of AI systems must be validated, as briefly mentioned in Chapter 2.2) and the **XAI Explanation Validation** (not mentioned). XAI explanations should be validated for both accuracy and alignment with human-interpretable features. Most XAI methods are post-hoc approximations (except for self-explainable neural networks), and their accuracy can be assessed using perturbation-based methods
(e.g., Goettling et al., 2024, https://doi.org/10.1038/s41598-024-63656-x; Lundberg et al., 2017, https://doi.org/10.48550/arXiv.1705.07874; Veerappa et al., 2022, https://doi.org/10.1016/j.jocs.2021.101539).

Furthermore, deep neural networks often learn non-linear feature combinations, so it is essential to verify the extent to which XAI explanations correspond to human-understandable concepts. While this is briefly mentioned in Chapter 5.2, the need for systematic validation
(e.g., as demonstrated by Hammer et al., 2025, https://doi.org/10.1038/s44387-025-00022-w) should be emphasized. Otherwise, “alignment with user expectations” (Chapter 6.2, (ii)) risks being conflated with expectation bias.

In Summary, the manuscript should stress the necessity of analyzing both the accuracy and the causal validity of explanations.

Methods

[3] The manuscript states that RQ1 addresses two main objectives, including the identification of design considerations for new or updated XAI systems. However, RQ1 as currently formulated (“Which factors influence end-user trust in AI systems?”) does not directly address the second objective. The research questions and objectives should be realigned to ensure comprehensive coverage and logical coherence.

[4] The use of the keyword “End user” in the literature search may have excluded context-specific studies that refer to user groups by name. As a result, the review may be biased toward more general or design-thinking-oriented articles. This limitation should be acknowledged. If, contrary to my understanding, various synonyms were used for end-user groups, this should be clearly communicated.

[5] The process by which the 29 trust-related factors were identified from the literature is not sufficiently detailed. It is unclear whether the process was purely inductive, deductive, or a combination of both, and whether individual factors were synthesized from multiple sources. The

methodology for factor identification and aggregation should be described in detail to enhance transparency and reproducibility.

Discussion

[6] While the identified factors are discussed, the manuscript lacks critical discussion of the reviewed articles themselves. It is unclear whether the analyzed studies provide a comprehensive overview of the field, and the domains represented (e.g., medicine, automotive, basic research) are not specified. Additionally, while the manuscript discusses its own limitations, it does not address potential limitations of the current state of research, such as gaps in coverage or methodological weaknesses in the analyzed studies. This discussion should be expanded.

Minor Comments
[7] Typographical issues: emrotional (Chapter 1, paragraph 3) > emotional; S. Kim, 2024 (Chapter 5.1, paragraph 1) > Kim, 2024
[8] Terminology: Replace 'candidates', 'candidate studies', etc. with 'articles' for consistency
[9] Does the study include papers published until March 2025 (Chapter 5, paragraph 1) or until January 2025, as previously described?

References
1. Goettling M, Hammer A, Malberg H, Schmidt M: xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. *Scientific Reports*. 2024; **14** (1). Publisher Full Text
2. Hammer A, Goettling M, Malberg H, Linke A, et al.: Fusion of automatically learned rhythm and morphology features matches diagnostic criteria and enhances AI explainability. *npj Artificial Intelligence*. 2025; **1** (1). Publisher Full Text
3. Veerappa M, Anneken M, Burkart N, Huber M: Validation of XAI explanations for multivariate time series classification in the maritime domain. *Journal of Computational Science*. 2022; **58**. Publisher Full Text
4. Lundberg, S., Acirc, M., Lee, S.-I. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30 (2017).

Are the rationale for, and objectives of, the Systematic Review clearly stated?
Yes

Are sufficient details of the methods and analysis provided to allow replication by others?
Yes

Is the statistical analysis and its interpretation appropriate?
Not applicable

Are the conclusions drawn adequately supported by the results presented in the review?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biosignal analysis, xAI, transparent AI

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Aug 2026

Iker Legarda

Dear Dr. Alexander Hammer, we sincerely thank you for your thorough, insightful, and highly constructive evaluation of our manuscript. Your feedback regarding technical validation, terminology clarity, search strategy limitations, and the critical appraisal of existing literature has been fundamental in elevating the scientific rigor and overall quality of our paper. Below, we provide a point-by-point response explaining how each of your recommendations has been addressed in the revised manuscript.

Comment 1: The terms transparency, explainability, and interpretability have distinct meanings within XAI literature, yet the manuscript sometimes uses them interchangeably and at other times does not. For clarity and academic rigor, these terms should be explicitly defined at the outset and used consistently throughout the paper.

Authors' response: We fully agree that establishing explicit definitions at the beginning of the manuscript is essential for conceptual clarity. We have added a dedicated paragraph in Section 2.2 to formally define transparency, interpretability, and explainability, as well as to describe their interrelationships: *"For an AI system to achieve meaningful explainability, it fundamentally requires both transparency and interpretability. Transparency refers to a system-level property reflecting the openness, auditability, and overall visibility of the AI pipeline, including its data sources, algorithmic architecture, and operational constraints (Arrieta et al., 2020). In contrast, interpretability represents an intrinsic structural property of a model, denoting the degree to which a human can directly comprehend its internal cause-and-effect decision logic without requiring auxiliary explanatory tools (Miller, 2019). Within this conceptual triad, explainability acts as the active, human-centered communicative bridge that synthesizes system transparency and model interpretability into contextually meaningful representations for non-expert users."*

Comment 2: [2] The technical validation of AI systems is a critical foundation for trustworthiness, yet this aspect is underrepresented in the manuscript. This includes the AI model validation (The accuracy of AI systems must be validated, as briefly mentioned in Chapter 2.2) and the XAI Explanation Validation (not mentioned). XAI explanations should be validated for both accuracy and alignment with human-interpretable features. Most XAI methods are post-hoc approximations (except for self-explainable neural networks), and their accuracy can be assessed using perturbation-based methods (e.g., Goettling et al., 2024, https://doi.org/10.1038/s41598-024-63656-x; Lundberg et al., 2017, https://doi.org/10.48550/arXiv.1705.07874; Veerappa et al., 2022, https://doi.org/10.1016/j.jocs.2021.101539). Furthermore, deep neural networks often learn non-linear feature combinations, so it is essential to verify the extent to which XAI explanations correspond to human-understandable concepts. While this is briefly

mentioned in Chapter 5.2, the need for systematic validation (e.g., as demonstrated by Hammer et al., 2025, https://doi.org/10.1038/s44387-025-00022-w) should be emphasized. Otherwise, "alignment with user expectations" (Chapter 6.2, (ii)) risks being conflated with expectation bias. In Summary, the manuscript should stress the necessity of analyzing both the accuracy and the causal validity of explanations.

Authors' response: We sincerely appreciate this critical insight. We agree that technical validation and causal validity are mandatory prerequisites for trustworthy XAI, as failing to validate post-hoc explanations risks creating an "illusion of transparency" or reinforcing expectation bias. We have incorporated this perspective into three distinct areas of the manuscript: 1. Theoretical Framework (Section 2.2): We introduced a explicit paragraph detailing explanation fidelity, perturbation-based validation, and causal grounding: *"Crucially, a foundational technical requirement for XAI is the validation of both model predictions and the explanations themselves. Since most XAI techniques rely on post-hoc approximations rather than self-explainable architectures, their explanatory fidelity must be rigorously assessed. Techniques such as perturbation-based methods (e.g., Goettling et al., 2024; Lundberg & Lee, 2017) are essential to verify whether an explanation accurately reflects the model's true decision mechanism. Furthermore, deep neural networks frequently capture non-linear feature combinations that do not naturally map to human reasoning. Therefore, systematic validation is required to ensure that explanations correspond to human-understandable concepts rather than ungrounded statistical artifacts (Hammer et al., 2025)".* 2. Categorization Taxonomy (Table 6): We updated the scope of the Evaluation Methods and Metrics category in Table 6: *"Addresses experimental protocols, including technical validation of post-hoc explanation accuracy and conceptual alignment with human reasoning, objective and subjective metrics, and iterative evaluation strategies with real users".* 3. Critical Appraisal & Discussion (Section 6.1): We expanded Section 6.1 to highlight the risk of conflating perceived transparency with true explanation accuracy, explicitly emphasizing that future research must analyze both technical accuracy and causal validity to foster genuinely justified trust rather than expectation bias.

Comment 3: The manuscript states that RQ1 addresses two main objectives, including the identification of design considerations for new or updated XAI systems. However, RQ1 as currently formulated ("Which factors influence end-user trust in AI systems?") does not directly address the second objective. The research questions and objectives should be realigned to ensure comprehensive coverage and logical coherence.

Authors' response: We have reworded the framing text in Section 3 so that each objective maps directly to its corresponding research question.

Comment 4: The use of the keyword "End user" in the literature search may have excluded context-specific studies that refer to user groups by name. As a result, the review may be biased toward more general or design-thinking-oriented articles. This limitation should be acknowledged. If, contrary to my understanding, various synonyms were used for end-user groups, this should be clearly communicated.

Authors' response: We agree that targeting general user terms rather than role-specific job titles presents a methodological constraint. We have expanded Section 7.3 (Limitations) to

transparently acknowledge this limitation: *"Secondly, prioritizing general user terms (e.g., "end-user", "non-expert") over role-specific titles (e.g., "clinician", "driver", "operator") may have omitted empirical studies referencing users solely by job designations who, despite domain expertise, remain AI non-experts. Consequently, the synthesis highlights cross-domain design principles rather than sector-specific operational requirements".*

Comment 5: The process by which the 29 trust-related factors were identified from the literature is not sufficiently detailed. It is unclear whether the process was purely inductive, deductive, or a combination of both, and whether individual factors were synthesized from multiple sources. The methodology for factor identification and aggregation should be described in detail to enhance transparency and reproducibility.

Authors' response: To clarify our methodological approach, we added Section 4.3 (Data coding and thematic synthesis) and Section 4.3.1 (Identification and aggregation of trust factors). These subsections explicitly explain that an inductive, bottom-up open coding process was executed independently by two researchers, followed by iterative grouping and consensus meetings: *"The identification of factors influencing non-expert trust followed an inductive, bottom-up open coding process. First, trust-related factors explicitly mentioned were recorded. These extracted factors were then iteratively grouped based on conceptual similarity. In a second analytical pass, factors cited in only a single article were re-examined to determine whether they could converge into broader thematic dimensions without losing their core meaning. Convergence was not artificially forced; unique factors were retained as standalone constructs whenever they represented distinct operational or psychological dimensions, such as System Complexity (present in 2 studies) and System Safety Assurance (present in 3 studies). This process yielded a final consolidated set of 29 trust factors."*

Comment 6: While the identified factors are discussed, the manuscript lacks critical discussion of the reviewed articles themselves. It is unclear whether the analyzed studies provide a comprehensive overview of the field, and the domains represented (e.g., medicine, automotive, basic research) are not specified. Additionally, while the manuscript discusses its own limitations, it does not address potential limitations of the current state of research, such as gaps in coverage or methodological weaknesses in the analyzed studies. This discussion should be expanded.

Authors' response: We have added a dedicated sub-section, Section 6.1 (Methodological landscape, domain distribution, and critical appraisal of the literature), to provide a critical appraisal of the analyzed literature.

Comment 7: Typographical issues: emrotional (Chapter 1, paragraph 3) > emotional; S. Kim, 2024 (Chapter 5.1, paragraph 1) > Kim, 2024

Authors' response: Thank you for pointing out these errors. They have been corrected as suggested.

Comment 8: Terminology: Replace 'candidates', 'candidate studies', etc. with 'articles' for consistency.

Authors' response: Thank you for pointing this out. All instances have been updated to "articles" throughout the manuscript.

Comment 9: Does the study include papers published until March 2025 (Chapter 5, paragraph 1) or until January 2025, as previously described?

Authors' response: We appreciate the opportunity to clarify this timeline. The primary database retrieval phase was conducted in January 2025, while the subsequent forward snowballing iterations retrieved 1 publication published in early 2025 (March 2025). To eliminate any potential confusion, we updated the text in Section 4.2.1 and Section 5: Instead of "Dropped to 1 in 2025 due to the closure of this review in March 2025", the text now reads "registered 1 publication in early 2025, retrieved during the forward snowballing iterations".

We hope that these comprehensive revisions and structural additions address all of your recommendations satisfactorily. Sincerely,

Competing Interests: No competing interests were disclosed.

Reviewer Report 29 April 2026

https://doi.org/10.21956/openreseurope.23842.r72826

© 2026 Olaniyi O. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Oluwaseun Oladeji Olaniyi
University of the Cumberlands, Williamsburg, Kentucky, USA

This manuscript presents a systematic literature review of the factors influencing trust in artificial intelligence among non-expert users, along with key design considerations for explainable AI systems. The study adopts a PRISMA-based approach combined with snowball sampling, using both backward and forward citation tracking to identify relevant studies. Beginning with an initial set of articles sourced from Web of Science, the authors apply structured inclusion and exclusion criteria and iterative screening processes to arrive at a final corpus of 35 studies. From this body of literature, they identify 29 distinct factors that shape user trust, spanning individual characteristics, sociocultural influences, system attributes, and interaction dynamics. In addition, they identify 120 design considerations for explainable AI, grouped into nine overarching domains, including explanation type, user experience, transparency, ethics, and evaluation.

The manuscript makes a clear and timely contribution by framing trust in AI as a multifactorial, dynamic, and context-dependent phenomenon. It emphasizes that explainability should not be treated merely as a technical feature, but rather as a relational and human-centered design

property. The discussion extends beyond synthesis to highlight emerging perspectives in the field, including the role of micro-level interaction cues, the evolving nature of trust over time, the importance of error handling and trust recovery, and the social dimensions of human-AI interaction. These insights position the work at the intersection of technical AI development and human-centered design, offering both theoretical and practical relevance.

The rationale for the review is clearly articulated, with the authors identifying a gap in understanding how non-expert users develop trust in AI systems and the lack of actionable design guidance for explainable AI. The objectives are explicitly defined through two research questions focused on identifying trust factors and extracting design considerations. This clarity of purpose strengthens the manuscript's coherence and provides a solid foundation for the analysis that follows.

The methodology is generally well-structured and transparent. The use of PRISMA guidelines, the justification for employing a snowballing approach, and the step-by-step description of the selection process all contribute to the credibility of the review. The inclusion and exclusion criteria are clearly stated, and the use of multiple reviewers in the screening process adds robustness. However, there are areas where methodological transparency could be improved. In particular, the search strategy relies on a relatively narrow set of keywords, which may have excluded relevant studies that use alternative terminology.

Additionally, reliance on a single database limits coverage breadth. While the snowballing method helps mitigate this to some extent, a more explicit discussion of these limitations would strengthen the methodological rigor. Furthermore, the process used to derive themes, such as how trust factors and design considerations were coded and synthesized, is not described in sufficient detail. Providing more information on whether the thematic analysis was inductive or theory-driven, and how consistency between reviewers was ensured, would enhance reproducibility.
The study does not employ statistical analysis, which is appropriate given its qualitative and interpretive nature. The use of thematic synthesis aligns well with the heterogeneity of the literature and allows for the integration of diverse perspectives across disciplines.

The conclusions drawn are largely supported by the results presented. The identification of trust factors and design considerations is consistent with the literature analyzed, and the argument that trust emerges from the interaction between user characteristics, system design, and contextual factors is well substantiated. The manuscript also demonstrates awareness of its limitations, such as inconsistencies in how trust is operationalized across studies and the challenges this poses for comparison.

However, some of the more forward-looking insights, particularly those related to emerging perspectives like micro-interactions and trust recovery, appear to be more interpretive extensions than direct outcomes of the reviewed data. While these contributions are valuable, the authors would benefit from clearly distinguishing between findings derived from the literature and their own conceptual extrapolations.

Another area that would benefit from further development is the relationship between the identified trust factors and the proposed design considerations. Although the manuscript discusses this linkage conceptually, it does not provide a systematic mapping or framework that

connects specific factors to corresponding design strategies. Introducing a structured model or visual representation would enhance the practical applicability of the findings and make the contribution more actionable for designers and practitioners.

To ensure the scientific robustness of the manuscript, several key issues should be addressed. First, the thematic synthesis process should be described in greater detail, including how categories were developed and validated.

Second, the limitations of the search strategy should be more explicitly acknowledged, particularly regarding keyword selection and database coverage. Third, the relationship between trust factors and design considerations should be more clearly articulated, ideally through a structured framework that demonstrates how the findings can be operationalized in practice.

In addition to these major points, there are some minor areas for improvement. These include ensuring consistency in reporting the number of identified trust factors, clarifying the rationale for their groupings, and providing more concrete examples of how the design considerations can be applied in real-world AI systems. Minor editorial refinements would also improve clarity and readability.

This manuscript represents a meaningful and well-executed contribution to the literature on explainable and trustworthy AI. It offers a comprehensive synthesis of a fragmented research area and advances an important perspective that integrates technical, ethical, and human-centered considerations. With relatively minor revisions to enhance methodological transparency and strengthen the linkage between findings and practical implications, the manuscript is suitable for indexing.

Are the rationale for, and objectives of, the Systematic Review clearly stated?

Yes

Are sufficient details of the methods and analysis provided to allow replication by others?

Yes

Is the statistical analysis and its interpretation appropriate?

Not applicable

Are the conclusions drawn adequately supported by the results presented in the review?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Aug 2026

Iker Legarda

Dear Dr. Oluwaseun Oladeji Olaniyi, we sincerely thank you for your thorough, constructive, and encouraging evaluation of our manuscript. Your insightful feedback regarding methodological transparency, the structure of our synthesis, and the practical mapping of our findings has significantly strengthened the rigor and overall quality of our paper. Below, we provide a point-by-point response outlining how each of your comments has been systematically addressed in the revised manuscript.

Comment 1: there are areas where methodological transparency could be improved. In particular, the search strategy relies on a relatively narrow set of keywords, which may have excluded relevant studies that use alternative terminology.

Authors' response: We agree that relying solely on initial keyword queries poses a risk of omitting relevant studies that employ non-standard or domain-specific terminology. To address this limitation and enhance methodological coverage, we implemented a systematic two-iteration snowballing strategy. We have explicitly clarified this procedure in Section 4.1.2 (Backward and forward snowballing in iterations): *To complement the initial database search and address potential limitations regarding keyword variations and index coverage, a two-iteration backward and forward snowballing approach was systematically applied, starting from the initial set of seven articles. This procedure mitigated the reliance on a single search engine and compensated for potential omissions of relevant studies that employed alternative terminology.*

Comment 2: Additionally, reliance on a single database limits coverage breadth. While the snowballing method helps mitigate this to some extent, a more explicit discussion of these limitations would strengthen the methodological rigor.

Authors' response: We have expanded Section 7.3 (*Limitations*) to transparently acknowledge the trade-offs involved in our search strategy, including database restrictions and search engine dependencies. The following text has been added to Section 7.3: *Secondly, specific methodological trade-offs in the literature search strategy must be explicitly recognized. The primary retrieval phase relied on a single database and a targeted set of keywords, which may have excluded relevant studies that employ alternative terminology or are indexed in other platforms. To mitigate these initial restrictions and compensate for potential search omissions, a systematic two-iteration backward and forward snowballing strategy was applied. However, this citation-dependent approach carries its own inherent limitations.*

Comment 3: To ensure the scientific robustness of the manuscript, several key issues should be addressed. First, the thematic synthesis process should be described in greater detail, including how categories were developed and validated.

Authors' response: We have added a dedicated sub-section, Section 4.3 (Data coding and thematic synthesis), which details the step-by-step process used to extract, code, and synthesize the data. Specifically, Section 4.3 clarifies that an inductive thematic synthesis approach was followed, involving two independent coders who conducted initial open coding, iterative axial coding to cluster codes into higher-order themes, and periodic consensus meetings to resolve discrepancies and ensure inter-coder reliability and validation.

Comment 4: Second, the limitations of the search strategy should be more explicitly acknowledged, particularly regarding keyword selection and database coverage.

Authors' response: As detailed in our responses to Comments 1 and 2, we have thoroughly integrated these methodological constraints into both Section 4.1.2 and Section 7.3 (Limitations).

Comment 5: However, some of the more forward-looking insights, particularly those related to emerging perspectives like micro-interactions and trust recovery, appear to be more interpretive extensions than direct outcomes of the reviewed data. While these contributions are valuable, the authors would benefit from clearly distinguishing between findings derived from the literature and their own conceptual extrapolations.

Authors' response: We have substantially extended the discussion in Section 6 with an operational mapping of the trust factors with the considerations for XAI design, and we have revised the framing text of subsection 6.3 (former 6.2) to explicitly inform the reader that these six emerging dimensions (e.g., micro-interactions, trust recovery mechanisms) represent synthesized, forward-looking conceptual extrapolations formulated by the authors based on patterns observed across the corpus, rather than direct textual findings quoted from individual articles.

Comment 6: Another area that would benefit from further development is the relationship between the identified trust factors and the proposed design considerations. Although the manuscript discusses this linkage conceptually, it does not provide a systematic mapping or framework that connects specific factors to corresponding design strategies. Introducing a structured model or visual representation would enhance the practical applicability of the findings and make the contribution more actionable for designers and practitioners. (Repeated in Summary: "Third, the relationship between trust factors and design considerations should be more clearly articulated, ideally through a structured framework that demonstrates how the findings can be operationalized in practice.")

Authors' response: To bridge the gap between our empirical findings (RQ1) and actionable guidelines (RQ2), we have developed Table 7 in Section 6. Table 7 serves as a structured mapping matrix that explicitly links each identified trust factor (e.g., Transparency, Predictability, User Control) directly to its corresponding design considerations and operational strategies. This visual framework provides a guide for practitioners seeking to translate theoretical trust factors into concrete interface and interaction designs.

Comment 7: There are some minor areas for improvement. These include ensuring consistency in reporting the number of identified trust factors, clarifying the rationale for their groupings, and providing more concrete examples of how the design considerations can be applied in real-world AI systems.

Authors' response: We have addressed all three minor points as follows: 1. Reporting

Consistency: We conducted a thorough audit across the manuscript to ensure numerical alignment. The total count of identified trust factors has been strictly standardized to 29 factors across the Abstract, primary text, and summary tables. 2. Grouping Rationale: We expanded Section 4.3 (Data coding and thematic synthesis) with two new subsections to explicitly detail our inductive methodologies. In particular, in Section 4.3.1 trust factors were aggregated based on conceptual similarity while preserving distinct psychological or operational constructs without forcing artificial convergence, whereas in section 4.3.2 we explain that design considerations were consolidated into 120 unique actionable guidelines to maintain operational granularity before being inductively mapped into nine functional domains. 3. Real-World Application Examples: In Section 6, we expanded the discussion of proposed design considerations to include concrete, domain-specific implementation examples (e.g., visual feature-attribution maps in clinical decision support, real-time boundary alerts in automated driving displays, and counterfactual simulation panels in financial risk assessment dashboards). We hope that these comprehensive revisions and structural additions address all of your recommendations satisfactorily. Sincerely,

Competing Interests: No competing interests were disclosed.