Review

Transparency and Explainability Focus: Making AI Decisions Interpretable to Humans

Aiperi Zhenishova *

Posted Date: 18 November 2025

doi: 10.20944/preprints202511.1366.v1

Keywords: transparency; explainability; automated decision-making; human-centered explanations; technology ethics

 doi:10.20944/preprints202511.1366.v1

Article

Transparency and Explainability Focus: Making AI Decisions Interpretable to Humans

Aiperi Zhenishova

Ala-Too International University, Kyrgyzstan; aiperi.zhenishova@alatoo.edu.kg

Abstract

This report presents a thorough consideration of the nascent area of Human-Centered Explainable Artificial Intelligence (XAI), concentrating on the crucial task of ensuring that AI decisions are understandable and credible to human users. With the spread of AI across sensitive domains like healthcare, finance, and online retail, the need for clear and understandable explanations increases. The review considers different formats of explanation such as visual aids (saliency maps, textual summaries) and analysis challenges faced in the evaluation process. Main finding: The research interest has also radically changed since 2021 from focusing on purely technical approaches to more on human perception, interaction and trust. We combine results of 73 published papers that exist until the year of 2024 from empirical research and show that local post-hoc explanation (particularly feature importance methods [e.g., LIME, SHAP]) is the current focus of much of the literature but that inherently interpretable models are treated with relatively little attention. Despite the large pool of explanation techniques, there is a dearth of standardized metrics to evaluate interpretability, user confidence, and impact on decision making. This gap restricts comparability of evidences between studies and hampers efforts to bring about efficient and user-friendly AI explanations. The paper calls for structured frameworks as well as a harmonized protocol for analyzing explainability - specifying how explanatory explanations would lead to greater user trust, understanding and support in making decisions. Ultimately, a humane, rigorous approach towards evaluating AI systems is necessary to not only make these transparent but also make them really understandable on the part of the reader. The goal of this work, in turn, is to drive further exploration to more trustworthy, human-centered modes of explanation that will bridge that the chasm between the complexity of algorithms and human understanding.

Keywords: transparency; explainability; automated decision-making; human-centered explanations; technology ethics

---

1. Introduction

As users increasingly require transparency of AI-driven systems, interpretable artificial intelligence (XAI) is being integrated in the technology field nowadays. It is essential to learn not only the decisions of artificial intelligence, but why they are made especially in sensitive fields like healthcare, banking, and online retail. Trust and informed decision-making is vital.

For example, there are many technologies that can help explain the decision-making by artificial intelligence (e.g., the use of charts, simple language summaries, interactive tools, and information to break down the thinking process of the model). It is generally not, however, clear which of these interpretation styles in fact makes it comprehensible to humans. At times, the information goes beyond the technical and becomes ambiguous or can't build user confidence without explanation, or difficult to get the use case for it. This illuminates the importance of solid mechanisms to track the extent to which these interpretations benefit the intended audience.

One of the primary challenges is that academic research and industry practice have taken up a plethora of evaluation approaches. Some users have an automatic scoring system that uses the AI itself using a set of pre-defined criteria for rating the quality of their interpretation. But some prefer manual

 doi:10.20944/preprints202511.1366.v1

feedback and inform user activity in questionnaires, action items, and by user interaction also with the system. It is crucial to figure out which evaluation techniques best represent the value of users and how they interpret those explanations.

An additional crucial aspect is to make evaluation criteria standardized. To compare the results together there must be alignment between the datasets of various AI projects to form shared benchmarks and indicators. These could be the speed at which an individual understands the reasoning, or how far trust it inspires, or if it makes the user get the job done better.

In my review, I looked through an abundance of recent work to chart the existing landscape of XAI evaluation. I described machine-based versus user-centered approaches and what interpretation to take. The best traits — like transparency, ease of use, trustworthiness, reliability, and support of user decision-making — are most often prioritized. My ultimate aim for my work is simply to push the evolution of artificial intelligence explanations that are genuinely user-friendly and relevant to individuals. As a result of outlining an unambiguous and thorough evaluation approach, we wish to assist developers in creating an AI system that is reliable and interpretable to users and rely on.

2. Methodology

2.1. Design of Study and Reasons

2.1.1. Selection of Research Format: Systematic Literature Review (SLR)

It was based on the fact that we were required a good and consistent approach for consolidating and critiquing all of the existing literature on how to assess Explainable Artificial intelligence (XAI). The XAI area is still new and fluid, so methods across the landscape are often not homogenous and well-matched - the SLR provides the best process to organize and critique such analysis. Applying a methodological framework will result in more objective, clear and logical interpretations of the outputs of previous studies and to contribute to a fuller understanding of the underlying literature in an open manner. By choosing, classifying and interpreting the various sources systematically, this methodology allows a clear reconstructuring of the current literature, as well as for the uncovering of recurrent theoretical contradictions and emerging paradigmatic patterns.

2.1.2. Safety of Objectivity and Reproducibility

We depend on the SLR structure to maintain objectivity and traceability in all aspects of the research, adhering to basic scientific reporting rules. We pre-crafted every detail — our search queries, our inclusion/exclusion rules, our categories of codes. This front-loading planning protects us from the selection bias common in comparatively free-form literature reviews, ensuring that we integrate XAI's diverse empirical data with confidence.

2.1.3. Analyzing Methodological Differences

As highlighted by [1], the evaluation studies of Explainable Artificial Intelligence (XAI) exhibit significant differences in both experimental designs and the assessment instruments employed. The Systematic Literature Review (SLR) methodology facilitates a thorough examination of this heterogeneity, not only by documenting current methodologies but also by enabling a critical assessment of the consistency and comparability of outcomes. This includes evaluating results derived from various methodological approaches, such as within-subjects versus between-subjects designs.

2.2. Define the Purpose and Conceptual Focus

The overarching objective is the systematic review and appraisal of empirical studies that are empirically focused on the Human-Centered Evaluation (HCE) of XAI system. The theoretical focus is aimed to shifting the focus from technical transparency to real human understanding, guided by three central conceptual pillars:

 doi:10.20944/preprints202511.1366.v1

2.2.1. The Human-Centered Property of Explainability

The methodology is closely drawn to the assertion that explainability is an inherently human-centered property [2]. Thus, the only limited inclusion criterion is a study that carry out empirical user evaluations involving human subjects. A careful omission of purely algorithmic transparency metrics ensures that the review touches the chasm between what the XAI is intended for (algorithmic output) and how it is understood by end-users (cognitive input).

2.2.2. The Shift from XAI to Explanatory AI (YAI) Paradigm

Another valuable theoretical focus of this study is examining the emerging shift from merely Explainable AI (XAI) to Explanatory AI (YAI). XAI is more focused on demonstrating how an algorithm works internally, but YAI is all about humanizing personally narrative-based reasoning, giving context to the user to aid in making sense of the decision. [3,4], present a coding framework that aims to methodically assess evaluation metrics according to their relevance to this conceptual shift. It distinguishes between metrics that primarily center on technical performance and metrics that assess how understandable or actionable the system's outputs are for human users. Thus, the analysis is designed to offer not only considerations of the technical capability of AI systems, but also the capability of those systems to communicate reasonable accounts to humans—a factor being recognized as central in the responsible and effective adoption of AI.

2.3. Research Questions

This systematic literature review aims to answer the following questions concerning the Human-Centered Evaluation of Explainable AI:

1. When did the primary interest in Human-Centered Evaluation (HCE) of XAI begin to rise ?
2. What is the main theoritical challenge this review addresses ?
3. What are the prevailing characteristics of Explainable AI (XAI) systems investigated within Human-Centered Evaluation(HCE) frameworks ?
4. Which methodological approaches and metrics are most frequently employed to evaluate the effectiveness of XAI from a user's perspective ?
5. What methodological and theoretical limitations (e.g., lack of standardization, external validity issues, cognitive underpinnings) hinder the consistent and comparable human-centered evaluation of XAI ?
6. Why is there a need to shift the focus from existing XAI evaluation metrics to a Human-Centered Evaluation (HCE) approach grounded in cognitive or social theories ?

2.4. Organizer Plan of Analysis and Review Stages.

We have summarised our complete breakdowns and then divided it into 4 stages that provide a basis. It provides we a framework by giving us a simple, method to use in organizing our own research.

I. Finding and selecting sources:

We will use Boolean search terms to refine our initial search terms; rigorous searching through an HCE criterion must be used to avoid biased results of choice.

II. Pulling Out and Tagging Data:

Our multi-layer Taxonomy of articles using their XAI properties, design, and metrics allows very complex findings to be chunked-into comparable and quantifiable elements.

III. Integrating Findings and Aggregating Them:

Thematic Synthesis takes this approach and interprets and ranks coded data toward a higher tier system to determine common behaviours and a best-practice taxonomy.

 doi:10.20944/preprints202511.1366.v1

IV. Appraising the Standard of the Literature:

This final stage reviews both the internal and external validity of reports, identifying common gaps where the research appears to be missing any theoretical underpinning to cognitive science or standardized measurement tools.

3. Search and Selection Strategy

The formulation of the search and selection strategy is a fundamental principle in the Systematic Literature Review (SLR) methodology, allowing both comprehensiveness (recall) and precision (relevance) of the retrieved literature set. This section outlines the existing methodological approach to database selection, and how to build a query, and the stepwise eligibility screening.

3.1. Definition of Data Sources and Formulation of Query

3.1.1. Selection of academic databases and search scope

The literature search was conducted in a number of top academic repositories to cover a broad range of published articles (Computer Science, Human-Computer Interaction (HCI), and Applied Artificial Intelligence) in a systematic and integrated manner. We included the following databases as our principal data sources:

ACM Digital Library and IEEE Xplore: These databases were chosen as critical sources for their prominent peer-reviewed papers from pertinent conferences (CHI, IUI, AAAI, et al.) and journals, with most empirical studies focused on XAI design and evaluation [2].

Scopus and Web of Science (WoS): These large use of citation indexing services have been used in scope to have wide disciplinary coverage and to retrieve journal articles with high impact from publications which might escape single-domain searches.

The list was time-based, to include all publications published from January 1, 2019, to December 31, 2024. This timeframe was selected precisely to describe the key point from 2019 onwards during which XAI scholarship transitioned quite rapidly from algorithmic transparency to human-centered assessment [5].

This means that not only could the collected literature be made more precise and relevant but also the methodologically sound synthesis of this material could be enabled to be provided and to perform as comprehensive an evaluation as possible on how much it engages with the technical and human aspects of XAI.

Search Query:

```
(Group 1:  "Explainable AI" OR XAI OR Interpretability OR Transparency)
AND
(Group 2:  "User Study" OR Empirical Evaluation OR "Human-Centered" OR Assessment)
AND
(Group 3:  Trust OR Comprehension OR Satisfaction OR Usability)
```

3.1.2. Boolean Search Query Formulation

A systematic Boolean search query was developted to identify unique relationships among three central notions: the XAI paradigm, the empirical frameworks used, and the objective human-friendly metrics for assessment. By covering all three at once, this comprehensive approach ensures that we can access resources that cover all three components only at the same time. The AND operator thus restricts search possibilities so only the searches that covered both the technical, as well as human aspects of XAI were recovered [6].

3.2. Inclusion and Exclusion Criteria (I/E)

We developed well-defined (I/E) inclusion (I) and exclusion (E) criteria and maintained a strict adherence to those during the screening process. To safeguard internal validity of our review and to avoid narrowing down, we only searched for articles that had included at least the following

 doi:10.20944/preprints202511.1366.v1

themes, with the hope that our review will demonstrate at least the basic tenets underlying human behaviour. Rigorously applying these rules ensures that the final collection of articles faithfully reflects the concepts and methods we set out to examine and reduces the risk for error stemming from poorly designed studies or inconsistent reports. Such strict application of set rules provides assurance that our final set of studies is an honest representation of the field without allowing for differences in study design differences or inconsistent reporting. Which enhances the overall reliability of the synthesis results obtained.

Table 1. Inclusion and Exclusion Criteria.

Category	Inclusion Criteria (I)	Exclusion Criteria (E)
I. Methodology	I1. The article should specifically mention an empirical user study involving data collection from human subjects [6].	E1. Works pertaining only to the XAI algorithms (e.g., technical advances to SHAP or LIME) that lacked human-subject validation [7].
II. Conceptual focus	I2. The key research question must center on evaluating the effect of XAI on cognitive or behavioral factors, such as Trust calibration, Perceived comprehension, or Decision-Making Quality [2].	E2. Secondary literature: e.g., narrative reviews, meta-analyses, theoretical essays, regulatory proposals, which would only be used for theoretical framing, never use in coding primarily data.
III. Source rigor	I3. Publications should originate from peer-reviewed journals or proceedings of high-tier academic conferences (e.g., IEEE Transactions, ACM/CHI Proceedings).	E3. Non-peer-reviewed media including preprints (e.g., arXiv), technical reports or workshop abstracts.
IV. Language	I4. Full-text publication should be English.	E4. Publications not in English.

3.3. Protocol: step-by-step process

We scanned the literature in a stepped screen fashion (3 steps) to ensure transparency and to avoid selection bias. This follows the best practices outlined in established guidelines, such as the PRISMA guidelines, to ensure the inclusion of a defensible final sets of studies in the field. This sequential, interrelated approach guarantees our final candidate studies are transparent and replicable. At this stage we can validate the methods and their supporting concepts in the process, then the synthesis is trustworthy.

Stage 1: Identification and duplicate removal

The raw pool of identified documents were original and collected as 1655 documents (see methods.pdf, Section 1). All the precise repeats removed prior to proceeding, to preserve the uniqueness of the bibliography.

Stage 2: Title and abstract screening

Remaining studies were screened in blinding for title and abstract.

Stage 3: Full-Text eligibility assessment

Full-text documents that had passed Stage 2 were retrieved. For each paper, rigorous reading of the Methodology and Results sections was performed in order to verify that all inclusion criteria (I1–I4) were met and to attend especially to the description of the experimental setup and participant recruitment.

 doi:10.20944/preprints202511.1366.v1

The final, curated set of literature, considered appropriate for the Data Extraction and Coding (Section 3) phase, consisted of 73 empirical studies [1].

4. Extraction and coding

The third phase of the Systematic Literature Review (SLR) includes Extraction and Coding of the data from the final set of 73 eligible empirical studies (identified in Section 2.3.) It is essential for the formation of a structured, quantitatively organized dataset that permits thematic synthesis and critical review of disparate literature.

4.1. Formulation of the coding and standardization protocol

4.1.1. Inter-Coder reliability (Reliability)

We standardised our data extraction form (DEF) to preprocess our data – common to systematic reviews and to prevent coders from interfering with their analyses. The DEF serves as a standard checklist in which all the main terms from each paper are recorded. Groups these into three major categories: How the XAI system functions; What experimental design was used in this study; and the evaluations and analysis procedures used for each case study. The article underwent two cycles of coding each in a sequenced manner, depending on each:

1. First-Cycle coding (Factual extraction): Direct factual data points were extracted (e.g., authors, publication year, specific XAI algorithm used: LIME/SHAP).

2. Categorical coding: Facts were arranged and classified (by pre-defined mutually exclusive categories) into the following categories, that will serve as the foundation of the review's analytical Taxonomy.

4.2. Coding characteristics of XAI systems (The Explanandum).

This subsection classifies information explaining to the user as the explanatory, the explanandum, and focuses particularly on whether the explanation takes the form of a substantive factor (what part of the reasoning of the AI is explained) or a modal factor (which aspect is the one that gets to communicate the information). These differentiations lend themselves to a sophisticated exegesis of the explanatory activity as a mediating link between algorithmic inference and human comprehension, which sheds light on modes of expression through which practices of explanation come to inform the interpretability and epistemic reach of decision-making in the light of AI. Based on their functional scope, XAI systems were classified on basis of the theoretical framework developed by [7]:

Local post-hoc explanations: Coding captured a particular system that was focused on a particular event (e.g., if one feature has significance for this one part of the system). These systems tackle the real-world issue: "Why was this particular prediction made?"

Global Post-hoc explanations: This coding encompassed systems showing systems that gave a global view of the model's performance over the entire dataset(s) - general behavior (e.g., general feature influence or decision boundaries). They answer the question: "How does the model typically work?"

Inherent interpretable models: Classification was used for naturally interpretable models: For models which have an interpretable architecture (linear models, decision trees, etc.), the architecture of the model gives the explanation (the model for understanding).

4.2.1. Format and type of data coding.

The classification specified the medium of communication and the AI task domain to be met, both of which affect the cognitive load of the user [2]:

Presentation format: Visual (e.g., Saliency Maps, feature plots), Textual (e.g., natural language sentences, rule sets), or Multi-modal (a combination thereof).

Data type and domain: Tabular data, image data (Computer Vision), or Textual data (Natural Language Processing).

 doi:10.20944/preprints202511.1366.v1

AI task: Categorization of the core machine learning task (e.g., classification, regression, recommendation).

4.3. Coding experimental design and stakeholders.

A descriptive report on implementation code of the study and researchers involved in the research is presented. This section provides, in detail, a complete account of the methodological method. These are characteristics of the participants involved.

4.3.1. Coding of experiment design scheme.

It was the design of the design scheme that was coded critically coding that determined the type of comparisons made as well as the description and the extent of comparisons made. internal validity of the results [6]:

Within-subjects design: Coded for experiments with exposure to all participants to treatment conditions (e.g., comparison of Explanation A vs. Explanation B).

Between-Subjects design: Coded for trials in which participants were randomized, and each group got only one kind of treatment (i.e., XAI vs. control).

Recruitment: Laboratory studies (highly supervised), online crowdsourcing (e.g., MTurk), or field studies (high ecological validity).

4.3.2. Coding of target audience (Stakeholders):

Participants were divided into a number of categories according to their role and expertise level since the XAI requirements are situationally variant [8]:

End-Users (Lay Users): Individuals that have no domain knowledge or domain expertise in Machine Learning.

Here we have grouped the evaluation metrics into two different buckets: subjective (what the User feels) and objective (what the User does). The key here is the analytical power to challenge how divergent evaluative paradigms operationalise “explainability”, and the extent to which these measures coalesce into a consistent methodological consensus.

4.4. Core evaluation metrics coding.

The evaluation approach to this review is the hardest part, as this addresses the actual metrics (dependent variables) used to demonstrate the efficacy of Explainable Artificial Intelligence (XAI). We’ve organized these evaluation metrics into two types: how the user feels (subjective) and what the user does (objective). Such differentiation provides the analytical leverage to ask just how widely diverse evaluative paradigms actually operationalize “explainability,” and to what degrees these measures converge toward a coherent methodological consensus.

4.4.1. Subjective measures

These metrics measure the user’s own perception and self-report, usually scored using Likert scales and questionnaires:

Trust: coded as perceived reliability of the system or the tendency of the user to depend on the AI recommendation. The specific Trust Scale or set of items used was also coded to measure metric variability.

Comprehension / Interpretability: Coded as the user interpretation, comprehensiveness, and perceived clarity Ma et al. (2024).

Satisfaction and Usability: Measurement of the overall user experience (UX) and interface effectiveness.

Cognitive load: the amount of effort that is required to process and interpret the provided explanation.

 doi:10.20944/preprints202511.1366.v1

4.4.2. Objective measures

These metrics track behavioral data as well as performance outputs, delivering measurable evidence of XAI effect:

User decision accuracy (UDA): It indicates if a person's decision at the end of the task is rooted in a reliable ground truth, in the form of an optimal outcome that could be applied in other circumstances (most often XAI or none at all).

Task completion time: This indicator also reflects efficiency measuring how fast users complete a task when given the explanation interface.

Faithfulness / Agreement: Indicates if a user's determination aligns with the inner logic of the model itself or if the user has to adjust their dependence, according to the agreed quality of explanation (Trust Calibration) [6].

5. Synthesis and Critical Appraisal

The third methodological phase involves the analysis and aggregation of methodological coding (Section 3) of that data to develop a conceptual framework and the identification of the critical methodological shortcomings based on existing evidence in Human-Centered Evaluations of XAI.

5.1. Synthesis process: thematic synthesis and taxonomy construction

5.1.1. Thematic synthesis approach

A thematic synthesis approach will be employed in order to correctly interpret the organized yet complex information gathered during the coding process of mosaic analysis. This qualitative approach involves an interpretation of emergent patterns and relationships, which breaks from a simple frequency analysis (typical of quantitative meta-analysis) for interpretation by qualitative synthesis researchers [5].

There are three steps in the synthesis process:

Code aggregation: The codes constructed in Section 3 (e.g., Local Explanation, Metric: Trust, Design: Within-Subjects) will ultimately become the initial codes.

Development of descriptive themes: These codes are combined in general, descriptive areas (e.g., "Prevalence of visual explanations," "Lack of objective performance metrics")

Generation of analytical themes: The descriptive themes generate new analytical themes, which are broad and general and serve as new taxonomy to characterize the relationships among XAI features, user groups and the evaluation results.

5.1.2. Construction of the conceptual framework

Main output of the synthesising will be a Conceptual Framework systematically mapping the design space of human-centred XAI evaluation [2]. This framework will help show the correlation of specific design decisions (e.g. explanation scope, communication format), to the human factors (e.g. Trust Calibration, Cognitive Load) measureme

5.2. The challenges of critical appraisal and standardization.

Critique, however, will require a clear and objective analysis of the quality, validity/rhetorical content of the literature reviewed that is not limited to summarization or summarization of result.

5.2.1. Metric diversity and failure to standardize.

An important aspect of the critical appraisal is an assessment of the lack of standardization across vital evaluative indicators, which curtails comparability between multiple studies:

The Trust Measurement Problem: This will explore how Trust and Comprehension are measured differently with unvalidated, ad-hoc scales. Given that subjective factors like Trust and Comprehension have been measured in many diverse ways, we can describe a methodological problem: the field has yet to define common criteria (a "gold standard") for measuring these critical XAI concepts [6].

 doi:10.20944/preprints202511.1366.v1

One other major evaluation assesses the external validity gap. The generalizability of low-stakes scenario and crowd-sourced (lay) users that allow for generalization to high-stakes domains (e.g. medical diagnosis or finance) are under scrutiny for external validity by [8].

5.2.2. Identification of theoretical and cognitive gaps

Identification of these gaps will identify theoretical and cognitive weaknesses. In this appraisal, we examine some enduring deficiencies in the empirical research on Explainable AI that hinder going from simple Explainable AI to truly effective Explanatory AI:

Cognitive Grounding Deficits The lack of literature discussing explicit grounding by cognitive and social theories (e.g., theories of explanation and cognitive load theory) on which XAI interface design is based will be probed. This deficiency is symbolic of a continuing emphasis on algorithmic transparency, in contrast with actual understanding by human beings [3].

The XAI vs. YAI Dichotomy: We make this distinction in this analysis through our use of the two-phasing separation between Explainable AI (XAI) and Explanatory AI (YAI). Here, this is a theoretical blind spot that leads to the problem that most of XAI communicates the model state from the inside, rather than making the pragmatic function of explanation as defined by [4].

5.3. Methodological limitations

As we discuss here, if transparency and academic integrity are to be maintained, the limitations and perils must already be pointed out since that is the order of the day that should exist as part and parcel of a Systematic Literature Review (SLR).

Publication Bias. The use of peer-reviewed academic databases carries the danger of publication bias, as studies that report statistically significant, or methodologically novel, findings are more likely to be published and indexed. This asymmetry possibly mischaracterizes the whole research environment, which contributes to overestimating XAI evaluation efficacy [1].

Language Restriction. However, the corpus limitation to English-language publications (criterion I4, Section 2.2) does mitigate the discrepancy and improve congruency of terminology in the corpus but decreases the corpus inclusion possibilities. As a result, articles issued in various language contexts may not be found and thus restrict the synthesis over different cultures.

Limited Access. However, while the search strategy has been executed with academic rigour across the multiple databases, limitation on access to specific indexed materials has resulted in pragmatic implications of corpus completeness. By not fully providing access to the full text, potentially useful studies were excluded, retaining a small but not insignificant sampling bias that may impact the interpretive potential of the review.

6. Results

Factual and systematic results are presented in this section, including the systematic findings of the literature review. The study was based on 73 empirical works about Human-Centered Evaluation (HCE) of Explainable Artificial Intelligence (XAI), published between 2019 and 2024.

6.1. Literature Descriptive Statistics

6.1.1. Publication dynamics and domain distribution.

The analysis indicated a significant increase in interest in HCE over recent years, similar to the work of Kim et al. (2024) by whose criteria almost all of the papers (about 93 %) started being published from 2021. This corresponds to a swift transformation in the XAI academia from algorithmic transparency per se to the concern on human perception and human interaction. Within application domains, HCE research methods are not uniformly distributed. Yao Rong et al. (2024) point out that XAI is still prevalent in a number of domains, such as recommender systems. At the same time, Kim et al. (2024) identified healthcare and education as the most commonly researched subject areas for this sample.

 doi:10.20944/preprints202511.1366.v1

6.1.2. Data Characteristics

The studies analyzed were dominated by 2 types of data:

Tabular as the most common type being due to its widespread use in various industries, i.e., business and finance.

Visual / Image data, in which Saliency Maps are commonly used for understanding computer vision models.

6.2. XAI System Features

XAI systems were structured along two main axes: the domain of explanation and the format of presentation which define what is being explained and how the intended information is delivered to the individual.

6.2.1. Explanation Scope

The overwhelming majority of the empirical research concerns local explainability:

Local Post-hoc Explanations: A common trend, involving approaches such as LIME and SHAP (which are used to provide explanations for a particular prediction), suggests the ease in answering a user's query: "Why did this decision happen?".

Global Post-hoc Explanations: Less studies on systems explaining the general behavior of the model across the entire dataset.

Inherently Interpretable Models: Models whose architecture is transparent (for example, linear models) were the least frequently studied.

[1] confirm this trend. In terms of explanation format, Feature Importance, a common kind of explanation that is often local, is the predominant explanation type (about 60%). The presentation format is visual (graphs, charts, saliency maps), often supplemented by textual (natural language) or rule-based explanations.

Design: No one standard is applied, and studies use both Between-Subjects and Within-Subjects designs. As noted by [6], this difference in design makes it not easy to compare the results for studies.

Recruitment: The major research technique adopted for recruitment was online crowd-sourcing (e.g., using MTurk), which had limited field studies with high ecological validity.

6.2.2. Target audience (stakeholders)

Most of HCE studies were performed on End-Users (Lay Users), those who are novice in machine learning. Such limitations pose a significant concern in terms of the external validity of results to high-risk domains (e.g., medical or financial experts) as described in the crucial Section 4.2.

6.3. Core Evaluation Metrics Coding

The effectiveness of XAI has been evaluated using subjective and objective measures. The analysis revealed a significant lack of standardization.

6.3.1. Subjective Metrics (User Perception)

The most commonly measured categories, according to the expectations methodology:

Comprehension / Interpretability: The most common metric (often expressed by the user as their self-assessment of how clear and complete the explanation is) [6].

Trust: The second most common metric.

The Measurement Problem: Importantly, there are not standardized, validated scales to measure Trust and Comprehension in common across studies, as was the case with the majority of studies. It backs up the conclusions of [1] on the "lack of standardization in methodology" and complicates results comparison between works [6].

 doi:10.20944/preprints202511.1366.v1

6.3.2. Objective Metrics (Behavior and Performance)

Objective metrics — measures of actual user behavior — were used far less frequently than subjective self-reports:

User Decision Accuracy (UDA): Estimate if the explanation results in an effective decision. **Faithfulness / Trust Calibration:** This metric is seen much less frequently, assessing how closely user dependence on the model matches actual model quality.

The review brings with it a striking finding in the imbalance between the large frequency of subjective metrics (e.g., feeling of comprehension) and a tiny number of objective metrics (e.g., true, objective decision accuracy)

6.4. Analysis and Synthesis of XAI Effectiveness

Combined review of the 73 studies indicated a lack of consistency in XAI effectiveness, a finding that is related to that of [6]:

Understanding: Explanations can be used to improve the users' subjective understanding process as to how a decision was reached.

Trust and usability: The impact of explanations on user trust and usability in building the trust and usability of the model is not sufficiently obvious and coherent [6]. Unjust or inappropriate trust calibration (over-reliance or undeserved distrust) is a common feature.

Fairness: Explanations haven't been particularly persuasive about ensuring that models make fair decisions [6]. Current XAI approaches are likely ineffectual in addressing the social and ethical elements of decision-making, which is further explained by the authors' findings.

7. Critical Appraisal and Discussion

But critique will require some hard-nosed, objective examination of the quality, validity and rhetorical content of the literature evaluated that is not confined to summarizing or summarising results.

7.1. Methodological Challenges and Standardization Gaps

An essential part of the critical appraisal, an important aspect of the critical appraisal, is the evaluation of the absence of standardization amongst significant evaluative indicators that limits relative comparability between various studies:

7.1.1. Variation in Metrics and Lack of Standardization

The Trust Measurement Problem explores the varying levels of trust and comprehension measurement by unvalidated (non-standardized) ad-hoc scales. As subjective dimensions such as trust and comprehension have been assessed in various varied ways, we can characterize a methodological issue: the field has not yet created common criteria (a "gold standard") for measuring such foundational XAI concepts [6].

The external validity gap (externally) is evaluated in another major evaluation. Generalizability in terms of low-stakes scenario, crowd-sourced (or lay users) users for high-stakes domains (e.g., medical diagnosis or finance) is also under examination for external validity [8].

7.2. Theoretical and Cognitive Gap (XAI vs. YAI)

Identification of these gaps highlights theoretical and cognitive weaknesses. In this evaluation, we consider some persistent limitations of empirical study of Explainable AI which prevent the transition from Explainable AI to really good Explanatory AI:

7.2.1. Cognitive Grounding Shortcomings

The absence of explicit grounding in cognitive and social theories (e.g., theories of explanation and cognitive load theory) from which XAI interface design is drawn to this end will be addressed.

 doi:10.20944/preprints202511.1366.v1

This limitation is also representative of the still apparent preference for algorithmic transparency over human understanding [3].

7.2.2. The XAI/YAI Dichotomy

In this analysis, we draw this contrast with our application of the two-phasing separation between Explainable AI (XAI) and Explanatory AI (YAI). In the present context, this is a theoretical blind spot which causes us to face the quandary that many of XAI communicates the state of a model from inside, instead of serving as an efficient explanatory function as described by Explanatory AI [4].

References

1. Kim, J.; Maathuis, H.; Sent, D. Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence* **2024**. Systematic Review, https://doi.org/10.3389/frai.2024.1456486.
2. Ma, S.; of Science, T.H.K.U.; Technology, C. Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies. *arXiv preprint* **2024**. Empirical Study Survey.
3. Christian Meske, Justin Brenne, E.S.; Dogangün, A. From Explainable to Explanatory Artificial Intelligence: Toward a New Paradigm for Human-Centered Explanations through Generative AI. *arXiv preprint* **2025**. Introduces the Explanatory AI (YAI) paradigm, emphasizing contextual reasoning, narrative communication, and adaptive personalization for human-centered explanations.
4. Sovrano, F.; Vitali, F. Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data. *arXiv preprint* **2024**. Introduces Explanatory AI (YAI) as a tool enhancing basic XAI output, grounded in Achinstein's theory of explanations emphasizing pragmatic, user-focused explanations.
5. Nguyen, T.; Canossa, A.; Zhu, J. How Human-Centered Explainable AI Interfaces Are Designed and Evaluated: A Systematic Survey. *arXiv preprint* **2024**. Systematic Review.
6. Yao Rong, Tobias Leemann, T.T.N.L.F.P.Q.V.U.T.S.G.K.; Kasneci, E. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**. Survey Paper.
7. Chinnaraju, A. Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. *World Journal of Advanced Engineering Technology and Sciences* **2025**. Review article.
8. Sam Baron, Andrew J. Latham, S.V. Explainable AI and stakes in medicine: A user study. *Artificial Intelligence* **2025**. User Study.