International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Explainable AI for Trustworthy Decision-Making: A Systematic Review of Explanation Quality, Evaluation Metrics, and Human Factors

Prof. Dr. Ramehshwar S. Pawar[1], Ms. Bhalerao Sanjeevanee Balaji[2]

[1]Principal, Swatantryasainik Suryabhanji Pawar Mahavidyalaya, Purna (Jn.) Dist. Parbhani
[2]Asst. Prof. Department of Library and Information Science, Swami Vivekanand Mahavidyalaya, Udgir Dist. Latur

ABSTRACT

Artificial Intelligence (AI) is increasingly used to support decision-making in different fields. However, understanding how AI systems produce their decisions remains a major challenge. Explainable Artificial Intelligence (XAI) aims to make these decisions easier to understand. This systematic review examines how the quality of AI explanations is evaluated, which metrics are commonly used, and how human factors influence the usefulness of explanations. The review focuses on technical measures, including fidelity, stability, consistency, and complexity, as well as human-centered factors such as understanding, trust, cognitive workload, and decision-making performance. Relevant studies will be identified through a structured search of scholarly databases and assessed using predefined selection criteria. The study-selection process will be reported in accordance with PRISMA 2020 guidelines. The review aims to identify common evaluation practices, examine the relationship between technical measures and human responses, and highlight gaps in existing research. The expected contribution is a structured understanding of XAI evaluation that can help researchers develop more comprehensive methods for assessing explanations and supporting trustworthy AI-based decisions.

Keywords: Explainable Artificial Intelligence, Trustworthy AI, Explanation Quality, Evaluation Metrics, Human Factors, Decision-Making, Systematic Literature Review.

1. Introduction

Artificial Intelligence (AI) is increasingly used in areas such as healthcare, education, banking, and agriculture. AI systems help analyze data, identify patterns, and support decision-making. However, many AI models provide results without clearly explaining how those results are obtained. This creates difficulties for users who need to understand the reasons behind an AI-based decision. Explainable Artificial Intelligence (XAI) has therefore gained attention as a way to make AI systems more understandable and transparent. XAI techniques provide different forms of explanation, such as identifying important features, presenting decision rules, or highlighting relevant parts of an image. These explanations can help users understand the factors that influence a model's predictions. However, the usefulness of an explanation depends on more than its technical accuracy. An explanation may represent a model correctly but still be difficult for users to understand or apply in practical situations. Researchers use several metrics to evaluate explanation quality, including fidelity, stability, consistency, and

complexity. In addition to these technical measures, human-centered evaluation examines whether users understand explanations, develop appropriate trust in AI systems, experience reasonable cognitive workload, and make informed decisions. Considering both perspectives is important because technically effective explanations do not always lead to better human understanding or decision-making. Despite the development of various XAI methods and evaluation measures, comparing their effectiveness remains challenging. Different studies may use different metrics, datasets, experimental settings, and user evaluation methods. This makes it difficult to determine which evaluation approaches are most suitable for different decision-making contexts. A systematic examination of existing research is therefore needed to understand current practices and identify areas requiring further investigation. The present study aims to systematically review research on XAI evaluation, with particular emphasis on explanation quality, evaluation metrics, and human factors. It examines the techniques used to assess explanations, the ways researchers evaluate human responses, and the relationship between technical evaluation measures and decision-making outcomes. The review also aims to identify methodological limitations and opportunities for future research. The study is expected to provide a structured overview of existing XAI evaluation approaches and highlight the importance of combining technical assessment with human-centered evaluation. Its findings may help researchers and practitioners develop more comprehensive evaluation methods for AI systems intended to support trustworthy decision-making.

2. Background and Theoretical Foundations

2.1 Explain ability and Interpretability

Explain ability and interpretability are important concepts in Artificial Intelligence (AI). They help users understand how an AI system produces its results. Although these terms are often used interchangeably, they may refer to different aspects of understanding an AI model. Interpretability refers to how easily a person can understand the internal behavior or predictions of a model. For example, a simple decision tree allows users to follow the rules used to reach a prediction. Explain ability refers to the methods used to provide reasons for a model's output. For example, an explanation method may identify the features that contributed most to a particular prediction.

Both concepts are important when AI systems are used in decision-making. Clear explanations can help users assess predictions, identify possible errors, and understand the limitations of a model. However, an explanation should not automatically be considered reliable simply because it appears understandable. Its quality and usefulness must also be evaluated. In this review, explain ability and interpretability are considered complementary concepts for understanding AI predictions and assessing their role in trustworthy decision-making.

2.2 Overview of XAI Techniques

Explainable Artificial Intelligence (XAI) uses various approaches to make AI-generated predictions easier to understand. These approaches differ in the way they explain a model's decisions. Some focus on a particular prediction, whereas others provide information about the model's overall functioning. LIME, introduced by Ribeiro et al. (2016), and SHAP, proposed by Lundberg and Lee (2017), are two widely used methods for explaining model predictions.

The main XAI techniques are described below.

1. Local Explanation Methods: These methods explain the factors influencing a specific prediction. LIME, for example, uses a simpler model to approximate the behavior of a complex model around a particular input.

2. Feature Attribution Methods: These methods estimate the contribution of individual input features to a model's prediction. SHAP is a well-known example of this approach.
3. Rule-Based Methods: These methods explain decisions through logical rules or conditions, allowing users to follow the steps that lead to a particular outcome.
4. Visual Explanation Methods: These approaches use visual elements, such as highlighted image regions, to indicate which parts of an input may influence a model's prediction.
5. Interpretable Models: Models such as decision trees and linear regression can be understood through their decision rules, structure, or coefficients.
6. Global Explanation Methods: These methods describe the general behavior of an AI model and help users understand how it makes predictions across different inputs.

Each technique has its own benefits and limitations. The choice of method depends on the model being explained, the application area, and the intended users. Therefore, evaluating XAI techniques requires attention to both the accuracy of the explanations and their practical usefulness. As discussed by Doshi-Velez and Kim (2017), interpretable machine learning requires systematic evaluation methods suited to the intended purpose.

2.3 Technical Evaluation Metrics

Technical evaluation metrics help researchers check the quality and reliability of explanations provided by AI models. Fidelity measures how closely an explanation matches the model's actual behavior. Stability checks whether an explanation remains similar when the input changes slightly. Consistency examines whether similar cases produce similar explanations. Complexity refers to how simple or difficult an explanation is to understand. Robustness checks whether an explanation remains reliable when the input contains small changes or noise. Completeness examines whether an explanation provides enough information to understand a prediction. Each metric evaluates a different aspect of explanation quality. However, a technically accurate explanation may still be difficult for users to understand. Therefore, researchers should consider both technical measures and human understanding when evaluating XAI systems (Doshi-Velez & Kim, 2017; Zhou et al., 2021).

2.4 Human Factors in XAI

Human factors are important in evaluating Explainable Artificial Intelligence (XAI) because explanations should help people understand AI decisions and use them correctly. Researchers consider factors such as user understanding, trust, cognitive workload, and decision-making performance. User understanding refers to how well people understand an AI-generated explanation. Trust concerns whether users rely on AI decisions appropriately. Cognitive workload refers to the mental effort required to understand an explanation. Decision-making performance examines whether explanations help users make better decisions. These factors may vary according to users' knowledge, experience, and the application area. An explanation that is technically correct may not be useful if users cannot understand it. Therefore, XAI evaluation should consider both the technical quality of explanations and their effects on human understanding and decision-making (Miller, 2019; Zhou et al., 2021).

2.5 Trustworthy AI and Decision-Making

Trustworthy Artificial Intelligence (AI) refers to AI systems that people can use with confidence because their decisions are reliable, understandable, and responsible. Explainable AI (XAI) supports this goal by helping users understand the reasons behind AI-generated predictions. In important areas such as healthcare, education, and banking, explanations can help users examine results and make informed decisions. However, explanations alone cannot guarantee that an AI system is trustworthy. The system

must also be reliable, fair, safe, and suitable for its intended purpose. Users should develop appropriate trust rather than accepting every AI prediction without question. Therefore, evaluating trustworthy AI requires examining both the technical performance of the system and the effects of its explanations on human understanding and decision-making (European Commission High-Level Expert Group on Artificial Intelligence, 2019; NIST, 2023).

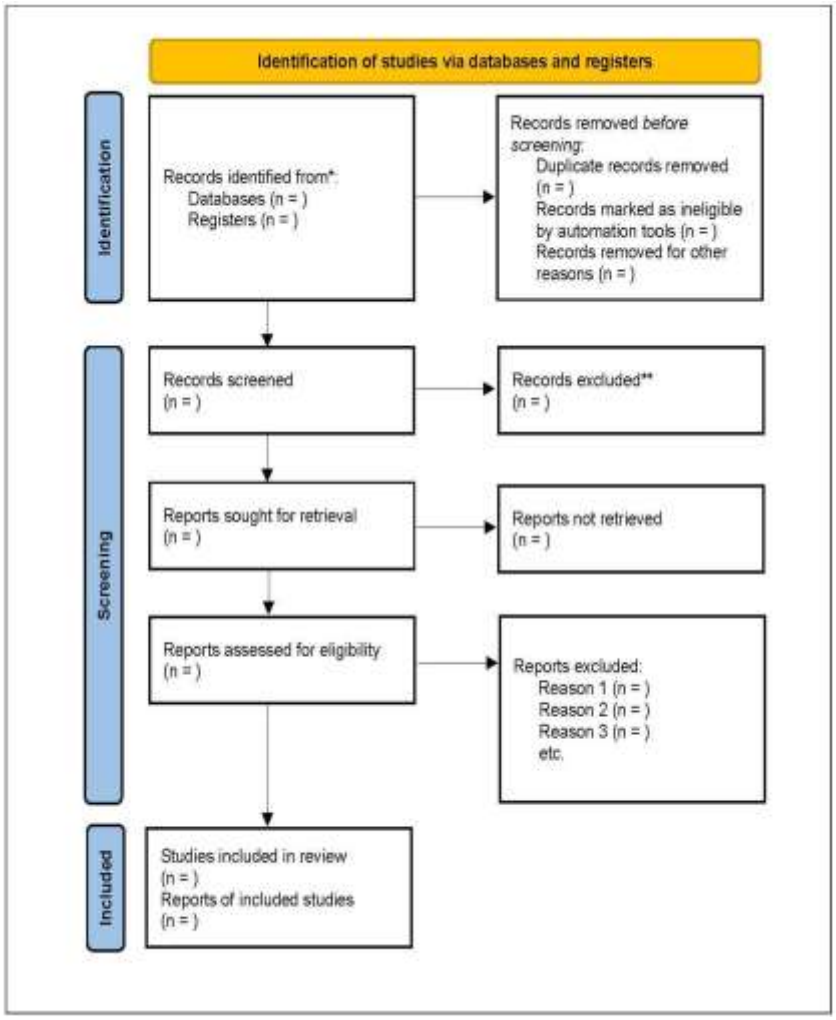


Figure 1: Proposed XAI Evaluation Framework

3. Research Methodology

3.1 Research Design

The study adopted a Systematic Literature Review (SLR) approach to examine Explainable Artificial Intelligence (XAI), focusing on explanation quality, evaluation metrics, and human factors. The review aimed to identify existing evaluation methods, compare their characteristics, and determine areas requiring further investigation. The reporting structure was based on the PRISMA 2020 guidelines (Page et al., 2021).

3.2 Databases and Search Strategy

The literature search covered Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library. Search terms included "Explainable Artificial Intelligence," "XAI," "explanation quality," "evaluation metrics,"

"human factors," "trust," and "decision-making." Boolean operators were used to combine these terms and identify relevant publications.

3.3 Inclusion and Exclusion Criteria

Studies were considered eligible if they were peer-reviewed journal articles or conference papers addressing XAI techniques, explanation evaluation, technical metrics, or human-centered assessment. Duplicate records, unrelated publications, non-scholarly documents, and studies that did not provide sufficient information relevant to the research questions were excluded.

3.4 Study Selection Process

The study-selection process involved identifying relevant records, removing duplicates, screening titles and abstracts, and assessing potentially eligible full-text articles. Each study was evaluated against the predefined inclusion and exclusion criteria. The selection process was organized according to the PRISMA 2020 reporting framework.

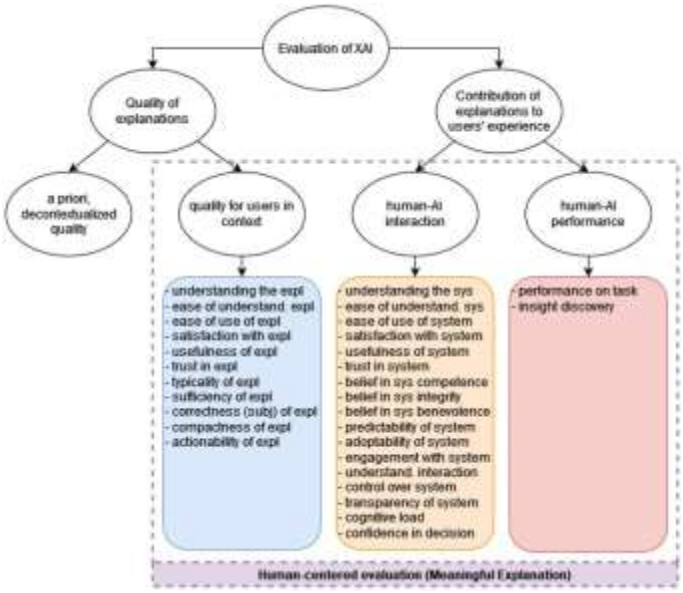


Figure 2: PRISMA Study Selection Flow Diagram

3.5 Data Extraction

A structured data-extraction format was used to organize information from the selected studies. The extracted information included publication details, research objectives, XAI techniques, evaluation metrics, application domains, human factors, principal findings, and reported limitations. This information provided a consistent basis for comparing the studies.

3.6 Quality Assessment

The methodological quality of the selected studies was examined using criteria related to research objectives, suitability of evaluation methods, clarity of experimental procedures, reporting of results, and acknowledgment of limitations. This assessment helped identify methodological strengths and weaknesses in the reviewed literature.

3.7 Data Analysis and Synthesis

The collected information was organized and analyzed through descriptive summaries and thematic synthesis. The studies were classified according to XAI techniques, technical evaluation metrics, and human-centered assessment factors. Particular attention was given to the relationship between technical

explanation quality and human outcomes, including understanding, trust, cognitive workload, and decision-making performance. The synthesis was intended to identify common evaluation practices, limitations, and opportunities for future research.

5. Results and Discussion

5.1 Overview of Existing Research

Existing research on Explainable Artificial Intelligence (XAI) shows increasing attention to the evaluation of explanations and their usefulness in decision-making. Earlier work focused mainly on developing explanation techniques and assessing their technical properties. More recent research has also examined how people understand, trust, and use AI-generated explanations.

Kim et al. (2024) reviewed 73 papers involving human-centered evaluation of XAI. Their review identified different ways of assessing whether explanations are meaningful to users. Similarly, Al-Ansari et al. (2024) reviewed 101 studies on user-centered XAI evaluation and examined evaluation approaches across different application domains. These reviews indicate that XAI evaluation involves both technical assessment and human-centered investigation.

5.2 XAI Techniques Used in Decision-Making

The literature describes several approaches for explaining AI predictions. Local explanation methods, such as LIME, explain the factors influencing an individual prediction. SHAP assigns contribution values to input features to help explain model outputs (Lundberg & Lee, 2017; Ribeiro et al., 2016). Other approaches include rule-based explanations, visual explanations, global explanations, and inherently interpretable models such as decision trees. The suitability of a technique depends on the model, the decision-making task, and the intended users. For example, technical users may need detailed feature contributions, whereas general users may prefer concise explanations in everyday language. Therefore, no single explanation technique can be assumed to be suitable for every application.

5.3 Evaluation Metrics for Explanation Quality

Zhou et al. (2021) reviewed methods and metrics for evaluating machine-learning explanations. Their work highlights the importance of assessing explanations through different quality dimensions rather than relying on a single measure. Common technical dimensions include fidelity, which examines how accurately an explanation represents the model; stability, which considers whether explanations change substantially under small input variations; and complexity, which concerns the simplicity of an explanation. Other measures examine consistency, completeness, and robustness. These measures address different properties and should be selected according to the purpose of the explanation.

However, technical metrics alone cannot establish whether an explanation is useful to a person. An explanation may accurately represent a model but still be confusing or unsuitable for a particular user. Consequently, technical measures should be considered alongside human-centered evaluation.

5.4 Human Factors in XAI Evaluation

Human-centered research examines how users perceive and interact with AI explanations. Important factors include understandability, perceived usefulness, trust, satisfaction, cognitive effort, and task performance. Kim et al. (2024) identified 30 components of meaningful explanations across the studies they reviewed. These components were organized into three broad areas: the quality of explanations in context, their contribution to human–AI interaction, and their contribution to human–AI performance. The review also found that trust and understanding were frequently evaluated, while approaches to measuring explanation quality varied considerably.

Al-Ansari et al. (2024) similarly reported that researchers evaluate XAI using different user-centered approaches and across several application domains. These findings suggest that the usefulness of an explanation depends not only on its technical characteristics but also on users' needs, knowledge, and decision-making context.

5.5 Relationship between Technical Quality and Human Factors

The reviewed literature indicates that technical quality and human-centered usefulness are related but distinct aspects of XAI evaluation. Fidelity or stability may show whether an explanation reflects a model's behavior, while user studies can examine whether people understand and use that explanation appropriately. Miller (2019) emphasized the importance of considering explanations from the perspective of the person receiving them. Similarly, Kim et al. (2024) distinguished the quality of an explanation from its contribution to human–AI interaction and performance. Together, these perspectives support a broader evaluation approach that combines technical measures with user understanding and decision-making outcomes. An explanation should not be considered successful merely because it increases users' confidence in an AI system. Evaluation should also examine whether users make appropriate decisions, recognize limitations, and avoid relying on incorrect AI recommendations.

5.6 Major Research Gaps

The literature identifies several challenges in evaluating XAI. First, researchers use different definitions, metrics, and experimental procedures, making comparisons between studies difficult. Kim et al. (2024) found that only 19 of the 73 papers they reviewed used an evaluation framework that had also been used by at least one other study in their sample. Second, technical evaluation and human-centered evaluation are not always integrated within the same study. Third, the usefulness of an explanation may vary according to the application, users' expertise, and the complexity of the decision. Finally, increased trust does not necessarily mean improved decision quality.

These challenges indicate the need for more consistent evaluation frameworks that combine technical reliability, human understanding, appropriate trust, and decision-making performance. Future research should also report evaluation procedures clearly so that findings can be compared and reproduced.

6. Conclusion and Future Scope

6.1 Conclusion

Explainable Artificial Intelligence (XAI) plays an important role in making AI-based decisions more understandable. However, the literature shows that evaluating explanations is a complex task because explanation quality involves both technical properties and human responses.

Zhou et al. (2021) highlighted the importance of evaluating explanations using different metrics, as no single measure is suitable for every explanation method. Similarly, Kim et al. (2024) found substantial variation in human-centered evaluation practices across the 73 studies included in their review. Al-Ansari et al. (2024) also examined user-centered evaluation approaches across 101 studies, demonstrating the broad range of methods used in this research area.

These reviews indicate that trustworthy XAI evaluation should consider technical characteristics, such as fidelity and simplicity, together with human factors, including understanding, appropriate trust, and decision-making performance. An explanation should not be considered effective merely because it is technically accurate or increases user confidence. Its usefulness should also be assessed in relation to the intended users and decision-making task.

Overall, existing research supports the need for comprehensive evaluation approaches that combine technical assessment with human-centered evaluation. Such approaches can help researchers better understand the strengths and limitations of XAI methods and develop explanations that support informed and responsible AI-assisted decisions.

6.2 Future Scope

Future research should focus on developing standardized evaluation frameworks that allow researchers to compare XAI methods more consistently. Technical measures and user-centered evaluation should be combined within the same studies wherever appropriate. Further investigation is also needed into how explanations affect user understanding, cognitive workload, appropriate trust, and decision-making accuracy in different application domains. Researchers should evaluate explanations with diverse user groups and realistic tasks rather than relying only on technical experiments.

Another important direction is the development of evaluation methods that consider the needs of different users, including domain experts and non-experts. Long-term studies could also examine whether explanations continue to support appropriate trust and effective decision-making after repeated interactions with AI systems. Finally, future studies should report their evaluation procedures, metrics, datasets, and limitations clearly. More consistent reporting and reproducible evaluation practices could improve the reliability of XAI research and contribute to the development of AI systems that support trustworthy decision-making.

References

1. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608
2. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
3. Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5), Article 593. https://doi.org/10.3390/electronics10050593
4. Al-Ansari, N., Al-Thani, D., & Al-Mansoori, R. S. (2024). User-centered evaluation of explainable artificial intelligence (XAI): A systematic literature review. Human Behavior and Emerging Technologies, 2024, Article 4628855. https://doi.org/10.1155/2024/4628855
5. Kim, J. K., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. Frontiers in Artificial Intelligence, 7, Article 1456486. https://doi.org/10.3389/frai.2024.1456486
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
7. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778