International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Explainable Artificial Intelligence (XAI): Building Transparency and Trust in Intelligent Systems

Samrudhi Parmeshwar Bhise
Department of Computer Science
Dr.D.Y Patil ACS college
Pimpri ,Pune ,India

Trupti Bajirao Kand
Department of Computer science
Dr.D.Y Patil ACS college
Pimpri ,Pune ,India

***Abstract*—Explainable Artificial Intelligence (XAI) is a emerging field of research focused on ensuring that AI decisions are transparent and interpretable, making them comprehensible to humans. As AI systems become increasingly complex and prevalent in high-stakes applications, XAI is crucial for fostering trust, accountability, and responsible AI use. This study offers a thorough review of XAI techniques, encompassing model-specific and model-agnostic methods, feature importance scores, and visualization techniques. We also underscore the significance of XAI in various application domains, such as healthcare, finance, and autonomous systems, highlighting its potential to boost AI adoption, identify biases and promote responsible AI development. Furthermore, we discuss future research directions and challenges in XAI, including the need for standardized evaluation metrics and human-centric design principles.**

Keywords*—Explainable AI (XAI) , Artificial Intelligence, Interpretability, Transparency, Accountability, Ethical AI, Model-specific techniques, Model-agnostic techniques, Feature importance scores, Visualizations, Healthcare, Finance, Autonomous systems

I) INTRODUCTION

Artificial Intelligence (AI) has progressed dramatically and has transformed from a group of experimental algorithms to a vital group of technologies within many different domains, including medicine, finance, and autonomous machines. Nevertheless, despite the dramatic opening of new and unprecedented possibilities provided by these and other AI technologies, there have been daunting difficulties in trust, accountability, and transparency. Conventional approaches to AI, and more notably deep learning technologies, function in a mysterious order and produce decisions that are difficult to grasp and interpret. Explainable Artificial Intelligence (XAI) has thus appeared as a timely answer to the aforementioned challenges. By enabling Artificial Intelligence to adequately explain the outputs it produces, XAI helps ensure the transparency of Artificial Intelligence and builds trust among its users. Experts have posited that trust in Artificial Intelligence does not, in its own right, represent a technical construct. Instead, trust represents a sociopsychological process that depends on the cognitive and perceptual properties of the user. Therefore, for Artificial intelligence to successfully provide explanations that resonate with humans, the explanations should match the cognitive and psychological properties linked to human factors and adapt to the context to be effective.

II) RESEARCH METHODOLOGY

This study follows a systematic literature review methodology to analyse and synthesize existing research on Explainable Artificial Intelligence (XAI) with, focusing on transparency and trust. Relevant research articles, conference papers, and academic reports were collected from widely used scholarly databases, such as Google Scholar, IEEE Xplore, SpringerLink, Elsevier ScienceDirect, and ACM Digital Library.

The literature search was conducted using keywords including *"Explainable AI," "XAI transparency," "AI trust," "model-agnostic explainability," "human-centred AI,"* and *"ethical AI."* Studies published primarily between 2020 and 2025 were considered to ensure the inclusion of recent advancements, whereas a few foundational studies were included for theoretical grounding.

Only peer-reviewed articles and reputed institutional publications written in English were selected for inclusion. Papers that lacked relevance to the explainability, transparency, or trust were excluded. The selected studies were analysed and categorized based on their focus on human trust models, explainability techniques, application domains, evaluation metrics and ethical considerations. This structured

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

approach ensured a comprehensive and unbiased review of the current state of XAI research.

III) LITERATURE REVIEW

A. *Human Factors in Trust Models*

Cheung and Shirley (2025) investigated the effectiveness of XAI in shaping human trust models. Their study emphasizes that explanations must align with cognitive processes, showing that trust is not merely a technical outcome but a socio- psychological construct. They argued that effective XAI enhances user confidence, particularly in critical decision-making framework.

B. *Trustworthy Applications of XAI*

Nasim et al. (2025) provide a broad survey of trustworthy XAI applications across biomedical engineering, data science, and information systems. Their work underscores the importance of domain-specific explainability, wherein trust is contingent on technical accuracy and ethical compliance. They also highlighted the challenges of scaling XAI across diverse industries.

C. *Reliable Metrics for Explainability*

The paper Bridging the Gaps in XAI stressed the need for standardized metrics to evaluate explainability. Without reliable measures, compliance and accountability are ambiguous. The authors propose frameworks for quantifying interpretability, fairness, and transparency, which are essential for regulatory compliance.

D. *Fostering Trust through Transparency*

Huma (University of Gujrat) examined the broader role of transparency in fostering trust in machine learning. The study concludes that explainability is not an isolated feature but is part of a holistic trust-building process involving governance, ethics, and user engagement.

E. *User Perceptions of Transparency*

Sunny (University of Maryland) explored how transparency influences user perceptions. The findings suggest that users value explanations that are concise, contextually relevant, and tailored to their level of expertise. Overly complex explanations can paradoxically reduce trust, highlighting the need for an adaptive and explain.

IV) COMPARATIVE ANALYSIS (Table Format)

Study / Author	Focus Area	Key Contribution	Limitation / Gap
Cheung & Shirley (2025)	Human factors in trust models	They showed that trust is sociopsychological and requires explanations aligned with cognition.	Limited to conceptual analysis; lacks empirical validation in diverse domains.
Sunny (University of Maryland)	User perceptions of transparency	Concise, context-aware explanations fostered greater trust than complex explanations.	However, it does not address scalability or domain-specific challenges.
Nasim et al. (2025)	Trustworthy applications across domains	XAI was surveyed in biomedical engineering, data science, and information systems, and domain-specific explainability was emphasized.	It highlights scalability issues and lacks a universal framework for cross-industry applications.
Bridging the Gaps in XAI	Reliable metrics for explainability	Proposed frameworks for standardized metrics (interpretability, fairness, transparency).	Frameworks remain theoretical, and no universally accepted metrics have been established.
Huma (University of Gujrat)	Transparency as holistic trust-building	Positioned explainability within governance, ethics, and user engagement.	It provides a broad perspective but lacks detailed technical methods for implementation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

V) EXPLAINABLE AI TECHNIQUES

Explainable Artificial Intelligence techniques aim to make AI model decisions understandable to human users. These techniques can be broadly classified into model-specific, model-agnostic, and post-hoc explanation methods, each serving different interpretability requirements.

A. Model-Specific Explainability Techniques

Model-specific techniques are inherently interpretable and designed for specific model architectures. Examples include decision trees, linear regression models, and rule-based systems, in which the decision-making process is transparent by design. These models allow users to directly trace how the input features influence the outcomes. However, their limitation lies in their reduced predictive power compared to complex deep learning models.

B. Model-Agnostic Explainability Techniques

Model-agnostic methods can be applied to any machine learning model, regardless of its structure. Popular approaches include Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive explanations (SHAP). These techniques explain predictions by locally approximating the model or assigning importance scores to features. Their flexibility makes them widely applicable, although computational complexity and stability issues remain challenging.

C. Global and Local Explanations

Global explanations describe the overall behaviour of a model, helping stakeholders to understand general decision patterns. In contrast, local explanations focus on individual predictions and explain why a specific output is generated. While global explanations are useful for model validation and governance, local explanations are critical in high-stakes decision-making.

D. Visualization-Based Techniques

Visualization techniques, such as feature importance plots, saliency maps, and partial dependence plots, enhance interpretability by providing visual insights into model behaviour. These methods are particularly effective for non-technical users, although they may oversimplify complex relationships.

VI) KEY FINDINGD/CHALLENGES AND GAPS

A. Patterns Identified

- Human-centric trust models: Studies have shown that trust in AI explanations is shaped by human cognition and psychology rather than purely technical constructs. Cheung and Shirley (2025) emphasize that effective XAI must resonate with human perception to build confidence.
- Transparency as a Driver of Trust: Transparency consistently emerges as the most critical factor in fostering user confidence. Sunny's findings highlight that concise and context-aware explanations are more effective than overly complex ones.
- Domain-Specific Applications: Nasim et al. (2025) demonstrated that XAI's effectiveness of XAI depends heavily on the domain. Healthcare requires clinical interpretability, finance demands regulatory compliance, and autonomous systems prioritize safety.
- Ethics and Governance Integration: Huma's research underscores that explainability is not isolated but is part of a broader ecosystem involving ethics, governance, and user engagement.

B. Gaps in Current Research

- Absence of Standardized Metrics: Despite proposals, there is no universally accepted framework for measuring interpretability, fairness, and transparency, leading to ambiguity in compliance and accountability issues.
- Scalability Limitations: Current XAI techniques often struggle to adapt across industries, with methods effective in biomedical engineering not directly transferable to finance or autonomous systems.
- Adaptive Explainability Deficit: Few practical implementations exist that tailor explanations to user expertise or context, despite their importance.
- Bias Mitigation Shortcomings: Although XAI can reveal biases, robust mechanisms for their active mitigation remain underdeveloped in the literature.
- Limited Empirical Validation: Much of the literature is conceptual, with fewer real-world case studies testing XAI in high-stakes environments.

VII) APPLICATION AREAS OF EXPLAINABLE AI

XAI plays a crucial role in multiple high-stakes domains where transparency and accountability are essential.

A. Healthcare

In healthcare, XAI assists clinicians in understanding AI-driven diagnosis and treatment recommendations. Explainable models help validate predictions related to disease detection, medical imaging, and patient risk assessment, thereby improving clinical trust and supporting ethical decision making.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

B. *Finance*

Financial institutions use XAI to explain decisions related to credit scoring, loan approval, and fraud detection. Transparency is essential for regulatory compliance and fairness, ensuring that automated decisions do not discriminate against individuals or groups of individuals.

C. *Autonomous Systems*

In autonomous vehicles and robotics, XAI enhances safety by providing insights into navigation, object detection, and control decisions. Explainability allows engineers and users to audit the system's behaviour, especially in critical or unexpected scenarios.

VIII) FUTURE RESEARCH DIRECTIONS

A. *Standardized Evaluation Frameworks*

Developing universally accepted metrics for interpretability, fairness, and transparency is essential for regulatory compliance and cross-domain adoption.

B. *Adaptive Explainability Models*

Research should focus on frameworks that dynamically adjust explanations based on user expertise, cognitive preferences, and contextual needs.

C. *Cross-Domain Scalability Studies*

Investigating how XAI techniques can be generalized or adapted across industries without losing effectiveness is crucial.

D. *Bias Detection and Mitigation*

Future studies must go beyond identifying biases to actively reduce them, ensuring ethical and equitable AI outcomes.

E. *Human-in-the-Loop Systems*

Incorporating user feedback into XAI design can enhance trust and ensure that the explanations resonate with diverse audiences.

F. *Empirical Case Studies*

More real-world experiments in healthcare, finance, and autonomous systems are needed to validate theoretical frameworks and measure practical impact.

IX) CONCLUSION

This study highlights the growing importance of Explainable Artificial Intelligence (XAI) in advancing transparency, accountability, and trust in AI systems. The literature review reveals clear patterns: trust in AI explanations is shaped by human cognitive and psychological factors, transparency is a critical driver of user confidence, and domain-specific applications demand tailored approaches to explainability. However, significant gaps remain, including the absence of standardized evaluation metrics, challenges in scalability across industries, and limited empirical validation in real-world contexts.

The implications of these findings are twofold. First, XAI must evolve beyond technical solutions to embrace human-centric and ethical dimensions, ensuring that explanations resonate with diverse users and contexts. Second, future research should prioritize the development of standardized frameworks, adaptive explainability models, and interdisciplinary collaborations to bridge gaps between theory and practice. By addressing these challenges, XAI can foster greater trust, promote responsible AI adoption, and ensure that AI systems contribute positively to society in healthcare, finance, autonomous systems, and other fields.

X) REFERENCES

1. Cheung, L., & Shirley, M. (2025). *Human factors in trust models: The role of explainability in AI adoption*. Journal of Artificial Intelligence Research, 48(2), 112–130.
2. Sunny, R. (2024). *Transparency and user perceptions in XAI*. University of Maryland, Department of Computer Science, USA.
3. Nasim A., Patel S., Wong J. (2025). *Trustworthy applications of explainable AI across domains*. International Journal of Data Science and Information Systems, 12(3), 215–240.
4. Khan, T., & Zhao, Y. (2024). *Bridging the gaps in XAI: Toward standardized metrics for interpretability and fairness*. Proceedings of the International Conference on Ethical Artificial Intelligence, 89–102.
5. Huma, S. (2025). *Transparency and governance in machine learning: A holistic trust-building approach*. University of Gujrat, Faculty of Information Systems.
6. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD Conference.
8. Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems (NeurIPS).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

9. Arrieta, A. B., et al. (2020). *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges*. Information Fusion, 58, 82–115.
10. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.