# Building Trustworthy AI: The Role of Explainability in Real World Systems

POOJA R.OZA, DR. SHRADDHA PHANSALKAR
Affiliation: Dept.Computer Enginegineering, MIT Art Design and Technology University,Pune
Email id: pooja.oza@mituniversity.edu.in, shraddha.phansalkar@mituniversity.edu.in

***Abstract*— In order to solve the lack of transparency and uninterpretability of advanced AI models, Explainable Artificial Intelligence (XAI) has become a vital area of study in AI. The importance of XAI in fostering openness and trust in AI systems is examined in this conceptual review. The work offers a cogent conceptual framework after analyzing the body of research on XAI and pointing out trends and gaps. To improve interpretability, a number of XAI strategies are covered, including rule-based explanations, saliency maps, attention mechanisms, and model-agnostic methods. The study looks at the problems that black-box AI models present, how XAI can improve transparency and trust, and how ethical issues and responsible XAI application should be handled. This review attempts to foster trust and accountable AI systems by encouraging transparency and interpretability**

***Keywords*— Explainable Artificial Intelligence, Black Box AI, Conceptual framework.**

### *Introduction*

Machines can now learn and solve problems thanks to artificial intelligence (AI), which has become a reality. AI systems evaluate data and get better at tasks through machine learning. From healthcare and finance to entertainment, this technology is quickly changing a number of industries. Artificial intelligence (AI) has many uses, ranging from self-driving cars to medical diagnosis, and its impact is only going to get bigger in the future.To ensure that AI is widely used, trust is necessary. In order for users to trust AI's accuracy and security, they must be aware of how it makes decisions (transparency), and have faith that it is impartial and fair. The gap between AI's potential and practical applications is filled by trustworthy systems. Explainable AI (XAI) tackles the challenge of trust in AI by making these powerful models more transparent. Unlike opaque "black box" models, XAI techniques reveal how AI arrives at decisions. This transparency fosters trust in three ways:

1) Users understand the reasoning behind AI outputs,
2) Biases in the data or algorithms can be identified and mitigated, and
3) Humans can monitor AI behavior and intervene when needed. XAI essentially bridges the gap between human understanding and AI capabilities, enabling responsible and widespread adoption of this technology.

With the increasing use of AI in practical applications, a key question has emerged: trust. The widespread adoption of AI is contingent upon is a crucial component of reliable AI systems, and this paper explores this topic. In order to build trust and enable the responsible integration of AI into a range of real-world applications, we will examine how XAI techniques contribute to transparency, fairness, and human oversight in AI.

### *I. UNDERSTANDING EXPLAINABLE AI*

In an effort to improve AI systems' interpretability and transparency, Explainable Artificial Intelligence (XAI) has become a significant area of study in AI. All of the XAI domains are arranged and their relationships to human users are shown in Figure 1.

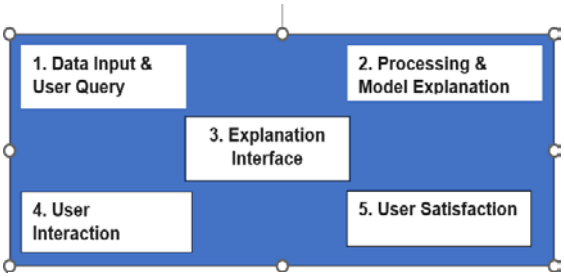


Figure 1. Review of XAI interaction with user

In order to shed light on the function of XAI for trust and transparency, this paper examines the body of prior research from a variety of fields and sources. The study by [2] focuses on the idea of causality and sets it apart from explainability. The authors contend that explainability relates to the system as a whole, whereas causability is an attribute of individuals. The basis for investigating the various facets and demands of explainability in AI systems is provided by this differentiation. [3] advocate for rigorous internal and external validation of AI models as a backup strategy in the event that appropriate explainability techniques are not available. They imply that more straightforward methods of accomplishing the objectives usually connected to explainability may be found in validation procedures. The authors advise against mandating explainability as a prerequisite for clinically applied models. [4] put forth a taxonomy to classify XAI techniques according to the levels of explanation, algorithmic approaches, and breadth of explanations. This

taxonomy aids in the development of self-explanatory, dependable, and interpretable deep learning models.[5] gave a thorough overview of the algorithmic ideas in XAI and offered insights into problems, prospects, and future developments. Their analysis AI System Information Realizing and Making Physical User Communications Interface for explanation of decision output Human in the loop: input Reasonable Artificial Intelligence Decision Justification The user questions the model Decision Justification For researchers and practitioners interested in the creation and use of XAI techniques, conveying a single explanation acts as a road map. Research that directly connect explainability to reinforcement learning (RL) models are evaluated in[6] in the context of RL. They shed light on various methods for achieving explainability in RL systems by classifying these studies into transparent algorithms and post-hoc explainability approaches. In light of developments in deep learning and machine learning, [7] offered an analytical assessment of the state-of-the-art in AI explainability. Their research sheds light on the issues that still need to be resolved as well as the advancements made. A changing picture of explainability in AI has been influenced by other significant works in the field of XAI, such as [8, 9], and [10]. Together, these studies deepen our understanding of XAI by analysing the cognitive processes that underlie explanation generation, putting forth taxonomies to classify XAI techniques, investigating different approaches to validation, and offering perspectives on the challenges and future directions of XAI.

## *II. Understanding Trust in AI*

XAI is essential to fostering openness and confidence in AI systems. This section addresses the role that XAI plays in building trust and looks at how XAI methods can reveal information about how AI models make decisions.

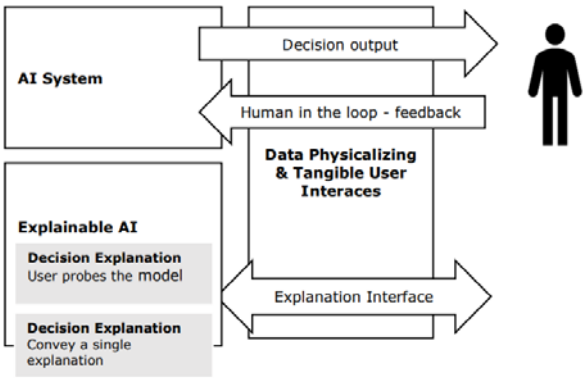


Figure2 : Understanding Trust in XAI

By giving users a better understanding of how AI models make decisions, XAI techniques boost user confidence in the system. Through XAI, users can better understand the underlying factors that the AI system takes into account by receiving explanations and justifications for the model's outputs [20]. This openness builds confidence and encourages faith in the technology.

The general transparency of AI systems is also enhanced by the insights offered by XAI techniques. Examining the decision-making procedures, spotting possible biases, and evaluating the accuracy and equity of the model's results are all possible for users. [21]

Transparent systems are essential for guaranteeing accountability and ethical considerations in domains where the consequences of AI decisions can have significant impacts, such as healthcare, finance, and autonomous vehicles.

Through the use of XAI techniques, users can better understand the inner workings of AI models and identify patterns, correlations, and potential errors or biases. Stakeholders are better equipped to make judgments, confirm that the model's predictions are accurate, and take necessary action as a result of the increased transparency [22].

In the field of medicine, XAI methods can help physicians comprehend the logic underlying AI- based diagnosis or treatment suggestions. Healthcare practitioners can improve patient care, make better decisions, and have more faith in AI-assisted healthcare systems with this understanding [23].

Comparably, XAI is necessary in the financial industry to guarantee accountability, equity, and transparency in AI systems. By providing insight into the elements influencing AI-based risk assessments and investment decisions, XAI techniques help regulators and investors verify the models' dependability and fairness, fostering greater accountability and trust [24]. By emphasizing important decision-pathways in the model and enabling human interpretation at different points in the decision-process, XAI is the transfer of understanding from AI models to end users [24].

### A. THE FACTORS THAT CONTRIBUTE TO TRUST IN AI

Using Explainable AI (XAI) Techniques to Increase Trust: Using XAI techniques can help increase trust in AI systems. XAI can: By elucidating the AI's reasoningprocess, it can Boost Transparency: People are aware of the "why" behind AI choices. Minimize Bias: XAI can assist in locating and resolving any biases that might exist in the data that the AI uses.Enhance user Experience: Confident communication with the AI system is the result of clear explanations.[30]

**A. Transparency:** People must comprehend how artificial intelligence (made possible by XAI) makes decisions. Uncertain systems breed mistrust.

**B. Fairness:** AI must be impartial and free from bias. Serious ethical repercussions may arise from biased AI.

**C. Accuracy and Reliability:** In order for AI to be trusted, it must produce results that are accurate and consistent.

**D. Robustness & Security**: To keep people's trust, AI systems need to be protected from manipulation and attacks.

**E. Privacy:** It's important to respect and safeguard user data privacy.

**F. Human Oversight:** Although AI automates certain tasks, human oversight is still necessary when making important decisions.

### *III. XAI TECHNIQUES FOR INTERPRETABILITY*

The inability to comprehend how AI systems and machine learning models process data and produce predictions or decisions is known as the "black box problem" in AI. These models frequently rely on complex algorithms that are difficult for people to understand, which undermines accountability and builds mistrust. Relationships between various industries and regulatory bodies have already been strained by the inability to effectively monitor and regulate AI systems [11].

A machine learning model's performance and its capacity to generate understandable and comprehensible predictions clearly trade off. In artificial intelligence, black-box models—such as deep learning and ensembles—are frequently employed, but they are difficult to understand [12].

Black-box models' lack of interpretability and transparency can result in dangerous decisions as well as grave errors. An attacker might, for example, alter the input data to sway the model's judgment and cause it to make risky or inaccurate decisions [13].

Thus, it is essential to create techniques that can aid in producing predictions that are easier to understand and interpret. Using self-explanatory interpretable models is one strategy. Using explainable AI is another strategy; this is still a very active field of study. Encouraging cooperation between various industries and regulatory agencies is becoming increasingly important in order to solve the black box issue.

The lack of adaptability, vulnerability to security holes, and difficulty in resolving issues with deep learning systems when they yield undesirable results are some of the drawbacks of black box AI models. Black box models can present a number of problems, such as unwelcome biases from the human world, even though they are appropriate in some situations [14].

It is crucial to create techniques that will enable the creation of predictions that are easier to understand and interpret. The application of interpretable models or explainable AI, which is still a work in progress, can help achieve this [15].

By solving the "black box" issue, we can encourage cooperation between various sectors and government agencies while guaranteeing the openness, accountability, and reliability of AI systems.

**Interpretable Models:**

*a. LIME*

Local Interpretable Model-Agnostic Explanations is referred to as **LIME**. This method is applied in Explainable AI to approximate the predictions of complex AI models in local domains, thereby rendering them interpretable [31][1].
In order for LIME to function, a black-box model's predictions for a given instance or sample must be translated into an interpretable model.
With the aid of this technique, users can comprehend why the AI model made a specific prediction for a given input. LIME sheds light on how AI models make decisions by concentrating on local explanations rather than requiring a comprehensive understanding of the model as a whole.[31]

*The function of LIME*

The process of developing interpretable models for comprehending AI predictions relies heavily on LIME, which stands for Local Interpretable Model-Agnostic Explanations [31]. It focuses on giving more transparent and intelligible explanations for each unique prediction generated by intricate AI models. LIME provides insights into decision-making processes by approximating the behavior of black-box models in particular local regions. Through the production of localized explanations, LIME enhances the interpretability and reliability of AI systems across a range of industries, including finance, law, and healthcare.

*b. SHAP*

**Additive explanations for Shapley (SHAP):**

Another method that is discussed in the paper is SHAP, which aims to explain individual predictions in a way that is consistent across the board [31]. By giving each feature a value of importance during the prediction process, it makes it easier for users to understand how the model behaves.

In order to make complex AI models more transparent and intelligible, LIME focuses on offering local explanations for individual predictions.By giving each feature an importance value for a specific prediction, SHAP, on the other hand, provides a game-theoretic method to explain the output of any machine learning model.[31]. Applications in a variety of fields, including law, finance, and healthcare, are made possible by LIME and SHAP, which both improve the interpretability of AI models [31].

### *IV The Ethics of AI in Action*

The application of XAI techniques and AI systems in general is heavily influenced by ethical considerations. In order to ensure responsible AI deployment, this section addresses the ethical implications of XAI.Concerns about accountability, justice, and bias in AI systems may be resolved by XAI. XAI can assist in identifying and reducing potential biases in algorithms or data by offering justifications for the decision-making process [25].

Fairer and more transparent decision-making processes are made possible by its ability to help developers and users understand how particular features or factors affect the results.

Maintaining accountability is yet another crucial component of responsible AI implementation. Through the use of XAI techniques, stakeholders can track the decision-making process and comprehend the elements that contributed to a specific result. This openness gives a way to assess the performance of AI systems and holds them responsible for their deeds [26].

Frameworks and guidelines have been proposed to encourage the moral and responsible use of XAI. The significance of transparency and explicability in automated decision-making processes is underscored by the General Data Protection Regulation (GDPR) of the European Union [27].

Fairness and explainability are also emphasized in the Responsible AI Principles, which were created by associations such as IEEE and ACM [28].

It is also necessary to take into account the wider societal impact when deploying XAI responsibly. This entails dealing with concerns about consent,

privacy, and possible social repercussions. It is imperative to develop and implement XAI in a way that upholds individual rights and advances societal well-being [29].

It is possible to promote the development of just, open, and accountable AI systems by incorporating ethical considerations into the design and application of XAI techniques.

### *V Results*

Interpretable AI models are created through the application of SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). The importance of transparent AI systems is emphasized as it covers the uses of Explainable AI (XAI) in healthcare, finance, and law. It is discussed how XAI has ethical and legal ramifications, emphasizing the necessity of fairness and accountability in AI decision-making processes.

In summary, the study emphasizes the value of XAI in guaranteeing that AI systems are reliable and intelligible in a range of contexts.

## CONCLUSION

The importance of Explainable Artificial Intelligence (XAI) in resolving issues with AI systems' transparency, interpretability, trustworthiness, and ethics is covered in the paper. It highlights how XAI methods provide insights into AI decision-making procedures, boosting confidence and empowering stakeholders to spot prejudices. Additionally, by facilitating accountability and adherence to ethical standards, XAI encourages the responsible deployment of AI. There are still issues to be resolved, such as managing complex models, harmonizing assessment techniques, and striking a balance between interpretability and performance. All things considered, incorporating XAI into AI development improves trust, tackles biases, and encourages ethical AI practices, opening the door for transparent, equitable, and reliable AI systems in the future.

## REFERENCES

[1] Colley, K. Väänänen and J. Häkkilä,, "Tangible Explainable AI - an Initial Conceptual Framework," in 21th International Conference on Mobile and Ubiquitous Multimedia, Lisbon, 2022.

[2] Holzinger, G. Langs and H. Denk, "Causability and explainabilty of artificial intelligence in medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, July 2019.

[3] M. Ghassemi, . O.-R. Luke and . A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, November 2021.

©ICS. Journal of Digital Art & Humanities, ISSN 2712- 8148, 4(1), June 2023 36

[4] G. Schwalbe and B. Finzel, "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts," Data Mining and Knowledge Discovery , 2021.

[5] J. Jiménez-Luna and F. Grisoni, "Drug discovery with explainable artificial intelligence," Nature Machine Intelligence, 2020.

[6] A. Heuillet, F. Couthouis and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," Knowledge-

Based Systems 214(7540):106685, 2020

[7] P. P. Angelov, E. A. Soares and R. Jiang, "Explainable artificial intelligence: an analytical Review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11(5), 2021.

[8] F. K. Došilović, M. Brčić and Nikica Hlupić, "Explainable artificial intelligence: A survey," in International Convention MIPRO, 2018.

[9] D. Gunning, M. Stefik and J. Choi, "XAI-Explainable artificial intelligence," Science Robotics, 2019.

[10] Michael Ridley, "Explainable Artificial Intelligence (XAI)," Information Technology and Libraries, 2022.

[11] S. Jagati, "AI's black box problem: Challenges and solutions for a transparent future," May 2023. [Online]. Available: https://cointelegraph.com/news/ai-s-black-box-problem- challenges and-solutions-for-a-transparent-future.

[12] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy (Basel), December 2020

[13] Kinza Yasar, "black box AI," March 2023. [Online]. Available: https://www.techtarget.com/whatis/definition/black-box-AI

[14] . L. Blouin, "AI's mysterious 'black box' problem, explained," 2023. [Online]. Available: https://umdearborn.edu/news/ais-mysterious-black-box- problem-explained.

[15] Rudin C., and . J. Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition," 2019. [Online].

[16] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising," 2013.

[17] .Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2014.

[18] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in the 22nd ACM SIGKDD International Conference, 2016.

[19] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 2017.

[20] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, 2018.

[21] B. Arrieta, N. D.-. Rodríguez and J. Del Ser, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion 58, 2019.

[22] R. Guidotti, A. Monreale and F. Turini, "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys , 2018.

[23] A. Rajkomar, E. Oren and K. Chen, "Scalable and accurate deep learning for electronic health records," Digital Medicine 1(1), 2018.

[24] Owens, B. Sheehan and M. Mullins, "Explainable Artificial Intelligence (XAI) in Insurance," Risks, 2022

[25] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM 61(10), 2016.

[26] J. Burrell, "How the machine 'thinks: Understanding opacity in machine learning algorithms," Big Data & Society 3(1,

January 2016

[27] Goodman and S. Flaxman, "EU regulations on algorithmic decision-making and a "right to explanation"," Ai Magazine 38(3), 2016.

[28] Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri and F. Turini, "Meaningful Explanations of Black Box AI Decision Systems," Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

[29] A. Jobin, M. Ienca and E. Vayena, "Artificial Intelligence: the global landscape of ethics guidelines," 2019.

[30] Arun Rai1 Explainable AI: from black box to glass box Arun Rai1 Journal of the Academy of Marketing Science (2020) 48:137–141

[31] G. P. Reddy and Y. V. P. Kumar, "Explainable AI (XAI): Explained," *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania,2023.

.