International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 7 (July 2026), PP 133-140

Explainable Artificial Intelligence for Trustworthy AI: Challenges, Solutions, and Future Directions

Muskan[1], Dr Brahmaleen Kaur Sidhu
[1,2]*Department of Computer Science and Engineering, Punjabi University, Patiala*
Corresponding Author: Muskan

ABSTRACT: *Explainable Artificial Intelligence (XAI) has emerged as a fundamental approach to improving the transparency, interpretability, and trustworthiness of artificial intelligence systems, particularly as complex machine learning and deep learning models become increasingly integrated into critical decision-making processes. Although these models achieve remarkable predictive performance, their black-box nature often limits users' ability to understand, validate, and trust their decisions. This paper presents a comprehensive analysis of the multidimensional challenges and limitations associated with existing XAI approaches, emphasizing their technical, ethical, and human-centric implications. The analysis explores key issues, including the accuracy–interpretability trade-off, explanation fragility, computational complexity, scalability, feature dependency, and the difficulty of generating explanations that are both faithful to the model and meaningful to diverse stakeholders. Furthermore, the paper examines ethical and governance challenges, such as fairness, accountability, regulatory compliance, and the risk of misleading or over-trusted explanations, which can significantly influence human decision-making. To address these limitations, the study discusses several pathways toward developing trustworthy XAI systems, including hybrid explainability techniques, standardized evaluation metrics, human-centered design principles, and emerging legal and regulatory frameworks. Real-world case studies from healthcare, financial services, and autonomous systems demonstrate the practical importance of explainability in supporting transparent, accountable, and reliable AI-assisted decision-making. The analysis highlights that achieving trustworthy AI requires an interdisciplinary approach that integrates technical innovation with ethical principles, regulatory governance, and user-centered design. The findings provide researchers, practitioners, and policymakers with a comprehensive understanding of current challenges, practical considerations, and future research directions for advancing transparent, trustworthy, and accountable artificial intelligence systems.*

Keywords: *Explainable Artificial Intelligence, Interpretability, Transparency, Trustworthy AI, Accuracy Interpretability Trade-off*

Date of Submission: 09-07-2026 Date of acceptance: 20-07-2026

I. INTRODUCTION

The extensive adoption of complex AI models, specifically deep neural networks, has revolutionized industries from defense to finance. However, their black box nature prevents humans from understanding their decision-making processes, which leads to significant concern regarding trust, accountability, and safety [6], [9]. Without a clear rationale, why does a model make a specific decision? Mainly because of this concern, end-users are restrained from adopting these technologies, especially in high-stakes environments [5], [11]. Explainable AI comes out as a direct response to this challenge, aiming to simplify these systems and provide insights into their internal working and outputs [1], [6]. This field seeks to bridge the gap between complex algorithmic decisions and human comprehension, thereby supporting humans' trust and enabling informed decision-making [4], [20]. This research paper provides a comprehensive, scholarly-level analysis of the limitations inherent in explainable artificial intelligence. The analysis suggests that these challenges are not solely technical but are deeply intertwined with ethical, societal, and human-centered factors. The paper then proposes a multi-faceted framework of solutions, supported by evidence from real-world applications, to navigate these complexities and build a trustworthy AI ecosystem. It is important to clarify that this analysis pertains to the technical and academic aspects of explainable artificial intelligence, not the corporate entity "XAI" founded by Elon Musk, which is an unrelated subject.

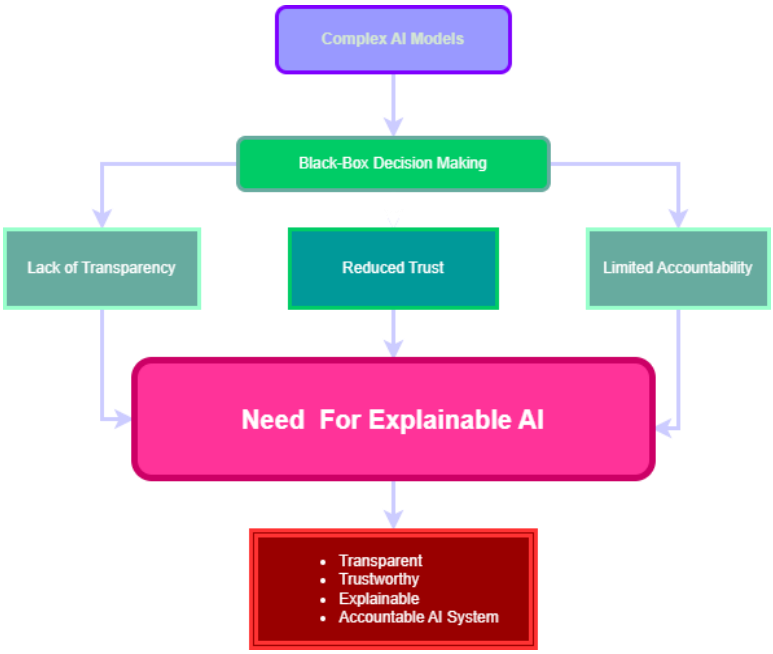

Figure 1: Evolution of Artificial Intelligence toward Explainable Artificial Intelligence

1.1 Objective of the study

1. To examine the fundamental concepts, principles, and techniques of Explainable Artificial Intelligence (XAI).
2. To analyze the major limitations of Explainable Artificial Intelligence, including the accuracy–interpretability trade-off, technical challenges, and ethical and human-centered issues.
3. To evaluate the role of existing XAI methods, such as SHAP, LIME, and Grad-CAM, in improving the transparency and interpretability of AI systems.
4. To investigate the significance of Explainable Artificial Intelligence in high-stakes domains, including healthcare, financial services, and autonomous systems.
5. To review existing approaches, regulatory considerations, and human-centered strategies for addressing the current challenges of Explainable Artificial Intelligence.
6. To identify future research directions for improving the transparency, reliability, and trustworthiness of Explainable Artificial Intelligence systems.

II. FOUNDATIONAL PRINCIPLES AND CONCEPTS OF XAI

Explainable AI is a field with a well-defined set of principles and methods designed to achieve its goals of transparency and trustworthiness [1], [9]. Core knowledge of these concepts is essential for a significant analysis of their limitations [6], [16]. While the terms are often used correspondingly, there is a key distinction between explainability and interpretability [4], [14]. Explainability focuses on providing a meaningful rationale for a specific decision, addressing the question, "Why did the model do this?" [2], [12]. This is often achieved through post-hoc methods that clarify a complex model's specific output. In contrast, interpretability aims to understand the overall behavior of a system by mapping input features to predictions, addressing the question, "How does the model work?" [4], [15]. Interpretability is often included in the model's design right from the beginning. Both concepts are critical for demystifying AI and building human confidence [1], [9].

Explainable Artificial Intelligence systems are based on several core principles that characterize a high-quality explanation. Those principles are [1], [13]:

Table 1: Core Principle of an XAI

Principle	Description
Explainable	These principles state that an AI system must provide evidence, Support, or reasoning for each decision made by the model.

Meaningful	This principle states that an AI model that explanation provided by the AI model must be understandable by, and meaningful to, its intended user.
Accuracy	The explanation must accurately reflect the system's process for generating the output.
Knowledge Limits	The system should operate only under conditions it was designed for and when it has sufficient confidence in its output.
Other Principles	Additional important principles include transparency, Justification, and robustness.

Beyond the principles outlined above, various methods in XAI can be broadly categorized along several dimensions [6], [8].

Table 2: Explainable Artificial Intelligence Methods

Methods	Description	Examples
Model-Agnostic	These techniques are specifically powerful because they can be applied to any black-box model, regardless of its internal architecture. This enables practitioners to interpret complex systems without understanding their inner workings.	LIME, SHAP
Model-Specific	Focus on explaining one particular type of model. They give a deeper understanding of that model, but it can't be used with different types of models.	Grad-CAM

Still further, explanations can be categorized as local and global. Local Explanation provides a rationale for a specific prediction or instance, such as why a particular loan application was rejected. Global Explanation provides a high-level overview of how the entire model makes its prediction, revealing general trends or the overall importance of features, such as the overall feature importance in a credit scoring model.

III. LITERATURE REVIEW

Recent advances in Explainable Artificial Intelligence (XAI) have focused on improving the transparency of machine learning models while maintaining predictive performance. A recent study on the accuracy–interpretability trade-off introduced a quantitative framework using the Composite Interpretability (CI) score to evaluate different machine learning models. The study demonstrated that highly accurate models generally exhibit lower interpretability, highlighting the inherent challenge of balancing model performance with explainability in practical applications [21].

The reliability of current XAI techniques has also been investigated in scientific domains. Research evaluating SHAP and LIME for process-level understanding in atmospheric sciences reported that these methods may generate misleading explanations when features are strongly correlated. Although the explanations accurately represent the model's behavior, they may fail to capture the underlying physical processes, limiting their reliability for scientific interpretation [18].

Several researchers have attempted to balance predictive performance with interpretability by combining different machine learning models. A hybrid framework integrating Bayesian Networks with Convolutional Neural Networks demonstrated that it is possible to achieve both high classification accuracy and meaningful explanations. However, the increased architectural complexity makes such systems more difficult to interpret and deploy efficiently [16].

Explainable AI has also shown promising applications in financial fraud detection. Recent studies employing SHAP and LIME demonstrated that explainable models improve transparency and assist investigators in understanding fraud prediction decisions while maintaining competitive detection performance.

Nevertheless, identifying complex fraud patterns and ensuring consistent explanations remain challenging tasks [6].

Another important concern is the trustworthiness of AI explanations. Studies on deceptive AI systems revealed that users often trust misleading explanations almost as much as truthful ones. These findings emphasize that persuasive explanations do not necessarily reflect correct model reasoning and highlight the need for rigorous evaluation of explanation quality before deployment [13].

To further improve fraud detection, stacking ensemble models combined with XAI techniques such as SHAP, LIME, Partial Dependence Plots (PDP), and Permutation Importance have achieved excellent predictive performance while providing interpretable explanations. Despite these advantages, ensemble architectures increase computational complexity, require longer training times, and present challenges for real-time implementation [9].

Comprehensive reviews of SHAP and LIME have further emphasized their importance as post-hoc explainability techniques capable of improving transparency for complex machine learning models. However, these methods are computationally expensive, and local explanations may become unstable when applied to datasets containing highly correlated features or noisy data [18].

Overall, the existing literature indicates that although significant progress has been made in improving AI explainability, several challenges remain unresolved, including the accuracy–interpretability trade-off, explanation reliability, computational scalability, human trust, and standardized evaluation. These limitations motivate the need for further research toward developing more trustworthy, reliable, and human-centered Explainable Artificial Intelligence systems.

IV. COMPREHENSIVE ANALYSIS OF THE MULTIDIMENSIONAL LIMITATIONS IN XAI

The field of XAI, despite its sententious promise, is confronted by a range of complex limitations that hinder its widespread and reliable implementation. These challenges are not merely technical but extend into the ethical and human dimensions of AI.

4.1 The Accuracy-Interpretability Trade-off: The underlying problem of XAI is the direct inverse relationship between the model's complexity, which often enhances the predictive accuracy and the interpretability of AI models. Highly complex or high-performance models like DNN or ensemble methods are generally opaque, making it difficult for humans to understand how they arrive at a result [6], [8]. In contrast to simpler or white box models like linear regression and decision trees, which are inherently interpretable in nature, but may underperform on tasks that need to capture complicated or non-linear relationships in data [4], [14]. This creates a fundamental tension: selecting a model that prioritizes one characteristic often means compromising on the other. This trade-off is not a universal constant for all cases, but is dictated by the specific use case and its associated risks. In a high-stakes domain like healthcare, where a misdiagnosis can lead to severe consequences, high accuracy is of the utmost importance. However, interpretability is also equally important for clinicians to trust the system, validate its findings, and explain its recommendations to patients. In such a situation, a slight fall in accuracy may be acceptable if the model's decision can be explained clearly and reviewed when it's needed. The need to balance these competing priorities has led to the development of a quantitative framework to measure and visualize this relationship, as demonstrated in a study on NLP models for predicting product ratings [21]. The following table, adapted from that research, illustrates this relationship by comparing models on their interpretability and classification performance accuracy (CPA) [21].

Table 4: Interpretability vs Accuracy

Model Type	Interpretability score (CI)	Classification Performance Accuracy (CPA)
VANDER	0.20	28.82
Logistic Regression	0.22	48.07
Naive Bayes	0.35	45.03
SVM	0.45	46.00
NN	0.57	49.00

BERT	1.00	53.82

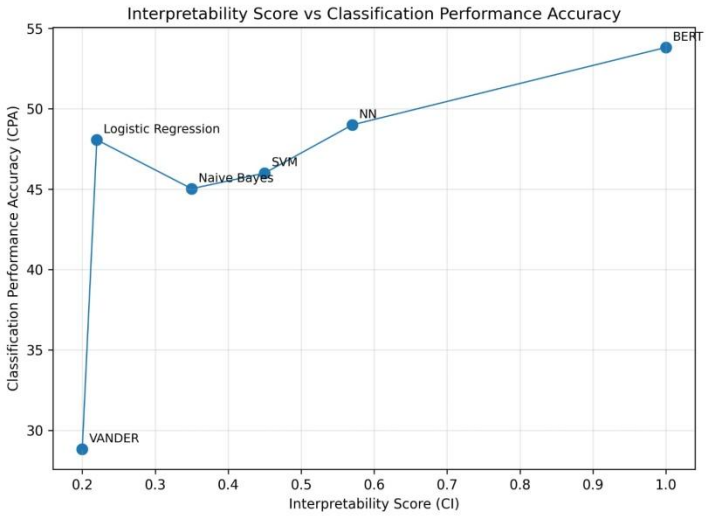

Figure 2: The Above data plotted in it. This study specific that a lower interpretability score means higher interpretability and shows that both characteristics are inversely proportional

The above table shows that as the model becomes less interpretable than their performance increases. This experimental observation provides measurable evidence rather than unscientific claims and highlights the key trade-off that XAI seeks to address [21].

4.2 Technical and Algorithmic Challenges: Apart from the above core trade-off, XAI encounters several technical and algorithmic challenges. One of the most significant problems is the issue of explanation fragility [8], [14]. Many XAI methods explain the AI model's behavior, but they don't explain the underlying real-world process or the dataset on which the AI model was trained [5], [17]. Because of this, the explanations can sometimes look correct according to the model, but still be misleading or wrong in real life, especially when the data has features that are strongly related to each other [8], [19]. Here, I consider an AI Model designed to predict students' academic performance. The model identifies both "hours studied" and "number of books owned" as correlated features influencing grades. Since students who dedicate more time to studying often also own more books, the model may generate an explanation suggesting that owning more books. Directly improve academic performance. However, in reality, the primary causal factor is the time spent studying, not book ownership. This reflects a spurious explanation that is consistent with the model's internal logic but misleading in a real-world context. A detailed analysis of a case study in atmospheric chemistry illustrates this point. The study shows that when XAI methods like SHAP and LIME are used on data where features are closely related, they can produce explanations that lead to an incorrect scientific understanding of the underlying process. These methods often assume that the features are independent, but this is often not true in real-world data. The causal chain is that the XAI method correctly explains the model's behavior based on false assumptions; although the resultant explanation accurately reflects the model, it is false in the real world [22]. A user relying on this explanation may derive a completely incorrect conclusion about the real physical or biological processes [9], [19]. This demonstrates that the faithfulness of the explanation towards the model doesn't guarantee its usefulness or correctness in the real world. Another major problem is computational overhead and scalability. Generating explanations, especially with post-hoc methods like SHAP and LIME, can be computationally intensive and difficult to scale for large datasets and real-world applications [2], [3]. This can make real-time explainability difficult to achieve, which is a significant barrier in time-critical applications like autonomous vehicles or real-time fraud detection [18].

4.3 Ethical and Human-Centric Hurdles: The limitations of XAI also go very far into the ethical and social aspects of AI. One such barrier is the user understanding gap. Even if the explanation is technically correct, it isn't useful or worthless if the end user can't understand it. Most of the AI tools are made by or for experts, so they are hard for non-experts, like business leaders or regular users, to use or understand. This makes it harder for these tools to be accepted and used, even if they work effectively. A deeper or major problem related to people is that they might trust AI explanations too much, which is

called Explainability Pitfalls (EPs) [23]. These are unexpected negative outcomes that happen when AI explanations are added, even if there's no intent to mislead the user. For example, users might depend too much on the system or trust it too much just because the explanation sounds believable, even if the decision is wrong. This is different from tricks meant to deceive users on purpose (called dark patterns) [5], [11], [23]. A study on large language models (LLMs) expressive that explanations those with misleading, deceptive reasoning can be more convincing than explanations that rely on honest reasoning and lead individuals or users to adopt disinformation. An explanation that sounds believable, when given to someone who does not understand the technical background details, can make them too confident or cause them to make a wrong choice. This can be harmful; it's like giving people a new medicine without testing or knowing about their disease. This shows why it's important to carefully test the medicine or know about the person's disease before giving them medicine. Likewise, this carefully tests XAI systems to make sure they are safe and helpful to people [18], [19].

V. A FRAMEWORK OF SOLUTIONS FOR A TRUSTWORTHY XAI ECOSYSTEM

In order to tackle the multidimensional limitations of XAI, there should be a comprehensive framework of solutions required, including technical, legal, ethical, and individual or human-centered strategies.

5.1 Technical and Methodological Solution: One of the most promising solutions for the accuracy and interpretability trade-off is the use of hybrid models [9], [21]. Basically, it means to combine both black-box and white-box model strengths together, where we get high performance and interpretability in one system [6], [20]. For example, a Deep learning model can be used as a feature extractor, and then its output fed into an interpretable model like linear regression for the final decision or output. A study on breast cancer diagnosis demonstrate the effectiveness of this strategy by using a hybrid of Bayesian networks and a convolutional neural network, which achieved both high accuracy and explainability [18], [19]. A critical gap in XAI research is a lack of standardization metrics to evaluate the quality, reliability, and faithfulness of explanations. Without such metrics, it is difficult to compare different XAI methods reliability which can interrupt the field's progress and adoption. Researchers are tackling this by developing new quantitative metrics. The Composite Interpretability (CI) [21] score is one new framework that measures interpretability by including various factors: simplicity, transparency, explainability, and model complexity, as indicated by the number of parameters. The method merges the expert assessments with quantitative measures to present an overall score and facilitate visualization of the trade-off between interpretability and accuracy, offering a much-needed underpinning for sound XAI analysis [21].

5.2 Legal, Regulatory, and Governance Solutions: The changing legal and regulatory environment is encouraging more use of explainable AI (XAI) [10], [11]. Novel or new rules increasingly require an AI system to be clear and explain how it makes decisions. This makes XAI an important tool to help organizations follow these regulations and stay compliant. For example, under the EU's General Data Protection Regulation (GDPR), individuals have the right to receive an explanation for decisions made by automated systems. The upcoming EU AI Act goes further by classifying certain AI applications, such as hiring and credit scoring, as "high risk". These systems will need details documentation of how they function and the data they use [10], [17]. In the US, XAI helps organizations comply with the FTC's guidelines and anti-discrimination laws by revealing and reducing hidden biases. Beyond compliance, XAI also strengthens accountability [5], [9]. By creating a clear record of how the decisions are made, XAI allows organizations to identify the source of an error, whether it comes from flawed data, biased algorithms, or system misconfiguration, and assign responsibility appropriately. In an area like autonomous vehicles, XAI is especially important. It can show the reasoning behind a machine's decision in the event of an accident, providing the transparency needed for legal review and responsibility [9], [20].

5.3 Human-Centered Solutions: Algorithmic transparency alone is not enough to build trust. A human-centered design approach is important to create explanations that are clear and useful for diverse groups. Who will use them? The ISO 9241-210 standard offers a four-step framework for HCD. It can be applied to explainable AI systems [24]:

1. Understand and describe the users and the context in which the AI will be used.
2. Define the needs and requirements of the users.
3. Develop design solutions based on these requirements.
4. Test and evaluate the design to ensure it meets the users' needs.

This method ensures that explanations are customized to fit the user's skills and knowledge, turning the AI from a "black box" into an interactive and clear system. Finally, the lack of trust in AI can be reduced through education and increasing understanding of governance[9], [11]. By helping teams and users learn how AI makes

decisions, trust grows, encouraging responsible use and adoption. A reliable AI system must be not only technically and legally correct but also designed for and understood by the people who use it [9], [[19].

VI. CASE STUDY: APPLYING XAI IN HIGH-STAKES DOMAINS

The theoretical challenges and proposed solutions for XAI are best understood through their application in high-stakes, real-world domains.

6.1 Healthcare: Enhancing Diagnostics and Clinical Trust: In healthcare, the black-box nature of AI models is a major barrier to adoption, where decisions can be a matter of someone's life or death [18], [19]. Clinicians require a clear understanding of an AI's rationale behind its decision to trust its recommendations or outcomes and to explain them to patients. To address this need, XAI techniques are implemented. For instance, in medical imaging, methods like Grad-CAM and saliency maps can generate visual heatmaps that highlight the specific areas of a scan that had the greatest influence on the decision [7], [20]. This provides a transparent, intuitive explanation that clinicians can use to validate the model's findings and gain confidence in its outcomes. A study on a COVID-19 classification system found that different stakeholders need different kinds of explanations, underscoring the importance of a human-centric design approach [18], [24]. Other research has shown that when explanations are integrated into clinical workflow, for example, by combining visual heatmaps with clinical data, clinicians' understanding and trust in AI-assisted diagnoses are significantly enhanced. The analysis of these cases confirms that XAI's role in healthcare is not just to provide an explanation but to build a bridge of trust between the clinician and the technology[9], [19].

6.2 Financial Services: Ensuring Fairness and Compliance: AI models used for credit scoring and loan approval need to be accurate, fair, and transparent to meet regulations and build customer trust [9], [19]. Baises inherent in historical data can cause unfair results, which are hard to detect without explainability [5], [17]. To address this problem, XAI methods like SHAP and LIME are used to explain how loan decisions are made. If a loan is denied, then the XAI system can clearly show which financial feature or parts of the credit history led to that decision. A new approach uses generative AI to create "applicant-friendly denial explanations", which give users clear and healthy feedback while helping banks meet regulatory requirements. XAI also plays a crucial role in detecting and preventing fraud [9], [17]. By explaining why, a transaction is marked as suspicious. XAI helps investigators focus on the most important cases and reduce false alarms [2], [6]. This improves efficiency and supports compliance with regulations. It also creates a clear audit trail, which is essential for meeting legal and regulatory standards [5], [17].

6.3 Autonomous Systems: The Road to Accountability and Safety: The safety and accountability of autonomous vehicles (AVs) are crucial [9], [11]. However, their complex systems often work as a black-box in nature, which makes it difficult to understand how they make decisions [6], [20]. This creates a lack of trust and slows down government approval. Explainable AI helps solve this problem by showing, in real time, how the vehicle's "thinking process" using visual displays and simple language explanations that people can understand [7], [20]. In case of an accident, XAI can give a clear and detailed record of how the vehicle made its decisions, which is very useful for insurance and legal investigations. XAI can also use "what-if" (counterfactual) explanations to show what the vehicle might have done in an alternative scenario, which helps in analyzing and fixing problems. For the perception of the system, XAI's techniques like Grand-CAM and saliency maps can highlight which objects, like people, other cars, or road signs, affect the vehicle's decision. This helps developers find weaknesses and correct system errors [9], [20]. However, it's important to note that XAI can only explain why a decision was made; it can't change the real-world limit of physics. One study showed that even nearly perfect autonomous vehicles may still face accidents that cannot be avoided due to timing or physical limits. In such cases, XAI's job is to explain why the vehicle acted as it did, ensuring transparency even when an accident couldn't be avoided [5], [17].

VII. DISCUSSION AND FUTURE SCOPE

The findings of this research paper highlight an important point: the challenges of the XAI are not just technical issues but are symptoms of a larger, socio-technical challenge. XAI is not merely a method for providing transparency; it is a critical enabling technology for achieving the broader goals of Responsible AI, including fairness, accountability, and safety.

Future research and development in this field must therefore shift from a singular focus on generating explanations to a more holistic, interdisciplinary approach that considers the entire XAI lifecycle. This includes:

1. Methodological Advancement: Creating more reliable and general explanation methods that can work with different models and data types. This also involves developing hybrid models that balance high performance with easy interpretability.
2. Standardization of Metrics: Establishing clear, standardized metrics to fairly compare and audit XAI systems. This is necessary for XAI to become a trusted, practical tool used in real-world industries.
3. Empirical Validation: Testing how XAI affects human understanding and decision-making through real-world studies. This helps avoid issues like misleading explanations and ensures that systems are designed with users in mind.
4. Interdisciplinary Collaboration: Encouraging teamwork between AI researchers, domain experts, ethicists, legal professionals, and end-users. Designing effective explanations is not just a technical task; it's a social process that requires input from many fields to build trustworthy systems.

VIII. CONCLUSION

The journey to build a truly trustworthy AI system is still a work in progress. XAI is a vital step toward this goal, but it also faces challenges such as the accuracy and interpretability trade-off and the danger of people trusting explanations too easily. The real strength of XAI isn't just in revealing how AI makes decisions, but linking those decisions to human understanding, laws, and ethics. By combining technical progress, strong governance, and human-centered design, we can turn AI from a mysterious system into a transparent and responsible tool that works together with people for the good of society.

REFERENCES

[1]. D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[2]. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[3]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[4]. C. Molnar, Interpretable Machine Learning, 2nd ed. Munich, Germany: Lulu Press, 2022.

[5]. B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2019, pp. 279–288, doi: 10.1145/3287560.3287574.

[6]. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[7]. W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," ITU Journal: ICT Discoveries, vol. 1, no. 1, pp. 1–10, 2017.

[8]. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2019, doi: 10.1145/3236009.

[9]. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[10]. European Commission, "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," Brussels, Belgium, 2021.

[11]. B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'," AI Magazine, vol. 38, no. 3, pp. 50–57, 2017, doi: 10.1609/aimag.v38i3.2741.

[12]. M. Rai, "Explainable AI: From Black Box to Glass Box," Journal of the Academy of Marketing Science, vol. 48, no. 1, pp. 137–141, 2020, doi: 10.1007/s11747-019-00710-5.

[13]. R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," arXiv:1812.04608, 2018.

[14]. Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.

[15]. D. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.

[16]. F. Dosilovic, M. Brcic, and N. Hlupic, "Explainable Artificial Intelligence: A Survey," in Proc. 41st Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210–215, doi: 10.23919/MIPRO.2018.8400040.

[17]. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.

[18]. E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[19]. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, 2019, doi: 10.1002/widm.1312.

[20]. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham, Switzerland: Springer, 2019.

[21]. P. Atrey, M. P. Brundage, M. Wu, and S. Dutta, "Demystifying the Accuracy-Interpretability Trade-Off: A Case Study of Inferring Ratings from Reviews," arXiv preprint arXiv:2503.07914, Mar. 2025, doi: 10.48550/arXiv.2503.07914

[22]. S. J. Silva and C. A. Keller, "Limitations of XAI Methods for Process-Level Understanding in the Atmospheric Sciences," Artificial Intelligence for the Earth Systems, vol. 3, no. 1, Jan. 2024, Art. no. e230045, doi: 10.1175/AIES-D-23-0045.1.

[23]. U. Ehsan and M. O. Riedl, "Explainability Pitfalls: Beyond Dark Patterns in Explainable AI," Patterns, vol. 5, no. 6, Art. no. 100971, Jun. 2024, doi: 10.1016/j.patter.2024.100971.

[24]. International Organization for Standardization, ISO 9241-210:2019, Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems. Geneva, Switzerland: ISO, 2019.