Toward Trustworthy AI: A Comprehensive Survey of XAI Methods, Software Toolkits, and Evaluation Standards

Simran[1], Anamika Rangra[2]

{simranjariyal71@gmail.com[1], rangra.anamika@gmail.com[2]}

Himachal Pradesh Technical University, Hamirpur, Himachal Pradesh, India[1,2]

Abstract. Explainable Artificial Intelligence (XAI) has emerged as a fundamental prerequisite for creating machine learning systems that are transparent, dependable, and accountable, particularly in fields where choices carry a high degree of risk. Despite the fact that numerous surveys have examined particular aspects of explainability, a more thorough approach is needed because XAI methodologies, software toolkits, evaluation protocols, and current synthesis are developing so quickly. Recent developments in three main areas are compiled in this review: (i) the methodological underpinnings of XAI, which include data level, intrinsically interpretable, and post-hoc explanation techniques; (ii) software and tool ecosystems that facilitate practical explainability through attribution based reasoning, visual analytics, surrogate modeling, and fairness or robustness assessment; and (iii) quantitative and qualitative evaluation standards that gauge explanation fidelity, stability, interpretability, human trust, and practical utility. This paper conducts a critical analysis of the advantages and disadvantages of the current approaches to XAI based on recent survey insights, benchmarking studies, and emerging evaluation frameworks. Highlighting the ongoing difficulties in implementing explainable systems in regulated, safety critical, and domain sensitive environments and fields.And also Standardized benchmarking, human-centered assessment, explanation robustness, and domain adaptive XAI in the review's conclusion as critical research areas that are crucial to the advancement of reliable and accountable AI.

Keywords: Explainable AI, Trustworthy AI, Interpretability, Post-hoc Methods, Evaluation Metrics, Transparency

1 Introduction

In the past ten years, artificial intelligence has advanced rapidly and has become a key component of many data driven decisions in a variety of high risk industries, including healthcare, finance, transportation, and governance. The systems are remarkably accurate and enable businesses to examine vast amounts of data that would be impossible to analyze manually. However, many of the models that support these advancements like transformer based designs, ensemble techniques, and deep neural networks are often opaque systems with complex internal mechanisms. Their behavior is the consequence of several interacting

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362522

elements, making it difficult to comprehend how a particular prediction is made. Concerns regarding justice, accountability, safety, and general regulatory compliance have consequently been raised by the quick increase in the use of AI in significant decision making. [1, 3]. The following example illustrates this problem more clearly. Imagine a digital loan screening tool that rejects an applicant and also does not provide any explanation for its decision. Even when the output is statistically sound, it is still essential for both the applicant and the lending institution to understand which factors led to the rejection—whether it is because of credit history, income patterns, or more troubling issues like hidden demographic correlations. When the reasons for a decision are obscure, users view the system as arbitrary and not trustworthy. Such situations illustrate why explanations are crucial whenever algorithmic decisions have an impact on people's lives. XAI thus developed as a response to these concerns and focused on techniques that would render model behavior explainable. Very early research concentrated on so called post-hoc interpretability: methods that studied trained black box models to identify relevant features or decision paths. Examples include surrogate models, saliency techniques, perturbation based analyses, counterfactual explanations, and attribution methods [2, 7]. While AI applications have diversified over time, the field of XAI has naturally kept pace with these changes. More recent contributions put forward taxonomies that include data level transparency, intrinsic interpretability during model design, multimodal explanation approaches [12], and end-to-end visibility along the entire machine learning pipeline.

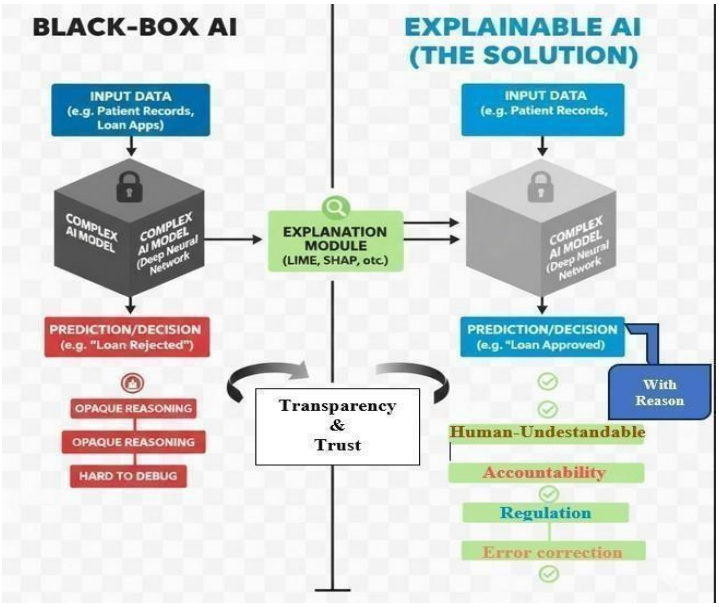


Fig 1. BlackBox vs XAI

With the recent emergence of LLM and multimodal foundation models, the scope of XAI has widened even further. New challenges regarding interpretability arise, such as explaining transformer attention patterns, prompt dependencies, and cross modal alignment in combined

models of text and images [13, 14]. Although such models unlock much power, hallucinations, latent biases, and unforeseen behaviors create challenges in these models, making the need to develop robust and transparent explanation techniques more critical every passing day. The difference between blackbox vs XAI is shown in Figure 1.

Apart from these theoretical developments, an expanding ecosystem of open source tools has made explainability practical. Local and global explanations, robustness assessment, fairness diagnostics, and human in the loop exploration are all made possible by libraries like SHAP, LIME, DALEX, Captum, InterpretML, AIX360, and OmniXAI. These tools are being incorporated more and more into practical workflows in a variety of industries, including public services, healthcare, finance, and retail.

But then also evaluating the quality and usefulness of explanations always remains a challenge. The effectiveness depends on factors such as fidelity, stability, comprehensibility, completeness, and robustness. Highly precise methods can overwhelm non-expert users whereas simpler approaches can skip important information. Different research groups use their own ways to evaluate explanations, which lead toward shared standards and stronger assessment methods.

Technical factors also affect explainability because different stakeholders have different needs. End users, regulators, domain experts, and data scientists etc. require explanations tailored to their objectives. In sensitive areas such as healthcare and finance, explanations must support professional judgment, satisfy legal and regulatory requirements, and enable safe, collaborative decision making [15, 17].

As evolve in transparency regulations, such as the EU AI Act, explainability has shifted from an optional to a mandatory requirement. In domains like clinical decision support, systems are expected not only to generate accurate outputs but also to provide interpretable reasoning that clinicians can understand and act upon. This discussion highlights the critical role of explainability in creating transparent, equitable, and dependable AI systems by summarizing the current state of knowledge, identifying key limitations, and outlining future research directions by integrating methodological, technical, and human centered perspectives.

2 Literature Survey

The field of XAI research has grown rapidly, which has been made possible by both technological advancements and demands to understand the behavior of automated systems used in real world scenarios. Much of the early work focused on justifying the need for explanations. The topics such as multimodal models, the inner workings of LLMs, human–AI interaction, and emerging policies for AI governance were discussed. This section outlines the significant contributions from three broad directions that help make sense of these developments: general surveys, methodological advancements, and studies focused on specific application areas.

Sajid Ali.[1] presented a comprehensive survey of Explainable Artificial Intelligence (XAI),which examined why explainability is essential given the opacity of modern deep learning models. It differentiates between white-box, gray-box, and black-box systems, and explains how XAI methods aim to improve transparency either through inherently interpretable

models or post-hoc techniques. This survey categorizes XAI approaches across data level, model level, post-hoc, and evaluation dimensions, and reviews available metrics, tools, and datasets. It also highlights legal, ethical, and user oriented drivers of explainability and highlights the need to adapt explanations for different stakeholders.

Ortigosa et al. [2] uses a practical perspective on Explainable Artificial Intelligence (XAI), illustrating how example based explanations, inherently interpretable model architectures, and feature attribution techniques such as LIME and SHAP can be integrated into real world workflows. The study underscores the importance of improving the transparency of complex machine learning models and improving their reliability and trustworthiness, thus supporting the interpretability and responsible use of automated decision making systems in practical applications.

Longo et al. [3] describe XAI more than a collection of algorithms in which XAI 2.0 manifesto tell that true interpretability must also account for how humans psychologically process explanations, along with legal obligations and ethical considerations.They bring up a simple but important point i.e rather than being the same for all end user classes, explanations should be tailored to each one. Developers, clinicians, auditors, and the general public all require different kinds of clarity. Thus, they appeal for research addressing personalization, robustness, and multi stakeholder transparency.

Several conceptual papers continue to influence how XAI is understood today. Pawar et al.[4] and Srihari[5] trace the journey from early rule driven systems to the complex neural networks in use today and offer readable explanations along the way. Xu et al.[6] put historical context on some of the earliest ideas shaping modern post-hoc interpretation.

More recent works have moved beyond simple feature attributions. Mathew et al.[7] discuss counterfactuals, interaction analyses, and hybrid approaches that combine visual and textual reasoning. These methods often mirror the ways in which people naturally ask “why this and not that?”, making them particularly suitable for real world scenarios. Domain specific studies add further detail. In healthcare, Rane et al.[8] demonstrate that transparent models enable clinical analysts to connect automated predictions to medically relevant indicators. In finance, C˘ernevi˘ciene˙ and Kaba˘sinskas[9] investigate transparency for credit scoring and fraud detection—processes where regulations require a clear justification of every automated decision. Kohan et al.[10] examine how interpretability tools support high stakes tasks such as the detection of intracranial hemorrhage.

Multimodal AI systems have generated their own set of explainability related issues. Rodis et al.[12] provide a summary of the developing domain of Multimodal XAI (MXAI), where interactions between text, images, audio and sensor input must be considered. Traditional single modality approaches often fail to capture these cross modal relationships. LLMs have pushed the interpretability frontier even further. Zhao et al.[13] review methods such as attention based analysis, activation tracking, and prompt focused interpretation. Since classic attribution methods often break down when applied to very large transformers, Ali et al.[15] propose improved propagation rules that more precisely highlight how these models form prediction.

Human–AI collaboration has become another central theme. Okonkwo et al.[14] argue that explanations are most useful when they help people and machines share responsibility in decision making, especially in high-stakes fields where trust and situational awareness matter.

Unlike static, one-off outputs, recent research highlights the importance of interactive explanations. Regulation is another critical concern. Hummel et al. [17] examines the relationship between XAI and the EU AI Act and demonstrate how openness helps with documentation, oversight, and risk management—tasks that modern AI systems are now expected to perform methodically.

Based on earlier tools, the discipline is still evolving. The concept of constructing small, local surrogate models—an approach that later inspired techniques such as SHAP and counterfactual explanations—was originally introduced with Ribeiro et al.'s LIME [11]. From these early conceptual contributions to a diverse ecosystem of algorithms, domain applications, human factors, and governance, the literature collectively presents a coherent narrative arc.

3 Taxonomy of XAI Methods

Explainable AI (XAI) includes many techniques that try to make machine learning models easier to understand. These methods can be grouped in different ways, such as by the kind of model they work with, the type of explanation they produce, or the purpose they serve. Having a clear taxonomy helps in comparing different approaches and choosing the most suitable one for a practical task. The main groups are listed in Table 1.

3.1 Intrinsic vs. Post-hoc Explainability

Some models are naturally easy to interpret. These are called intrinsic methods. Their structure is simple enough for a person to follow, so we can see how the model reaches a particular output. Decision trees, linear models, rule based systems, and additive models fall into this group. They are transparent, but they may not perform well on very complex problems.

Other explanations are created after a complex model has already been trained. These are known as post-hoc methods. Instead of opening the model, they analyse its behavior from the outside. They usually rely on feature importance scores, local approximations, counterfactual examples, or visual tools like saliency maps. LIME and SHAP are common examples. These methods are useful when accuracy is important, but we still need some level of interpretability.

3.2 Global vs. Local Explanations

XAI methods can also be divided based on whether they explain the whole model or just one prediction:

- **Global explanations** describe the overall behavior of the model. They help us understand which features matter the most on average and are useful for auditing and documentation.

- **Local explanations** focus on one particular input. They answer specific questions such as "Why was this applicant rejected?" and help build trust at the individual level.

Both types are important because they address different needs.

3.3 Model-Specific vs. Model-Agnostic Approaches

Some techniques are tied to a particular model architecture. These are called model- specific methods. For example, convolutional neural networks have activation maps, and transformers use attention based interpretation.
Other methods work with any model, regardless of how it is built. These are model-agnostic methods. They treat the model like a black box and study only the inputs and outputs. LIME, SHAP, partial dependence plots, and many counterfactual explanation tools fit in this category.

3.4 Main Categories of XAI Techniques

Several broad types of XAI methods are commonly used:

- **Feature Attribution:** Shows how much each feature contributed to a prediction.
- **Surrogate Models:** These methods use a simpler, more interpretable model—such as a small decision tree or linear model—to approximate the behavior of a more complex, often opaque model. By doing this, they help academics and practitioners better comprehend and trust the complicated model's outputs by offering insights into how it produces predictions.
- **Explanations Based on Examples:** offers comparable models, prototypes, or significant examples.
- **Counterfactual Explanations:** The minimal changes to input features that would change the model's decision.It basically helps users to understand model behavior.
- **Rule-Based Methods:** It represents knowledge and decision logic using rules specifically if-then rules, which can be interpreted directly by humans that are simple and carry transparency in decision making.
- **Concept Based Explanations:** These approaches explain AI predictions using high level, human understandable concepts instead of raw features. For example, in image classification, a model's decision can be described using concepts like "round shapes" or "striped patterns," making it easier for non-technical users to understand the reasoning.
- **Visualization Techniques:** It Highlight areas of the input that influence model predictions using tools like attention maps, heatmaps, or saliency maps, help users to quickly understand which parts of an image, text, or data drive the model's decisions widely used in computer vision and NLP.

3.5 Causal Explainability

The goal of traditional XAI techniques is to highlight the features that influence a model's predictions. Causal XAI aims to identify the causes behind an outcome. It provides deeper insight into model behavior by analyzing how changes in one variable propagate through the system, using tools such as do-calculus, causal graphs, and counterfactual reasoning.

- **Causal graphs:** A causal graph is a structured diagram that shows how one variable can affect or lead to changes in another. It helps us see the cause-and-effect relationships in a model in a clear and simple way.
- **Counterfactual queries:** Counterfactual queries examine how the model's output would change if one specific input were altered, while everything else stays the same. This helps

us understand “what would happen if things were different.”

- **Do-calculus:** Do-calculus is a set of rules used in causal XAI to study how deliberate interventions on certain variables influence the final outcome. It helps identify how changing one factor directly affects predictions.

These methods are helpful in places like hospitals, courts, and banks. They help us see why the model gives an answer, so people can trust it and make better choices.

3.6 Illustrative Example

Here is a small example that shows how different XAI methods can explain a loan model in different ways.

- A feature attribution method like SHAP identifies credit score and income as the most influential factors in the decision.
- A local surrogate model, such as a small decision tree used to represent the reasoning behind a single prediction.
- A counterfactual explanation might show how a loan decision might be different if the applicant’s income were marginally greater.
- Stronger dependencies can be shown by a causal explanation, which demonstrates how income and loan approval results are impacted by job stability.

Each method provides a different perspective and helps different people understand the model according to their needs.

Proposed XAI Taxonomy. The framework illustrated in Figure 2 integrates four key concepts: data transparency, model transparency, post-hoc explanations, and explanation evaluation. Together, these components offer a practical way to understand how different XAI approaches interact within the broader system.

4. Regulatory and Compliance Perspectives

The introduction of regulations such as the EU AI Act shows that there is a growing need for transparency, clear documentation, and continuous human oversight in high risk AI systems. Therefore, Explainable AI (XAI) helps to address these requirements by making model decisions more transparent, easier to understand, and easier to audit.

Table 1: Simple Taxonomy of XAI Method

Category	Description
Intrinsic Methods	Transparent models such as decision trees, rule based systems, and linear models.
Post-hoc Techniques	Methods that generate explanations for complex black box models after training. Includes tools like LIME and SHAP.
Global Explanations	Describe overall model behavior across the entire dataset; useful for auditing and documentation.

Local Explanations	Focus on explaining individual predictions to help users understand the reasoning behind a specific output.
Model-Specific Methods	Techniques tailored to specific architectures (e.g., CNN activation maps, transformer attention).
Model-Agnostic Methods	Methods applicable to any model by analyzing inputs and outputs (e.g., LIME, SHAP).
Feature Attribution	Shows how much each input feature contributed to a model's prediction.
Surrogate Models	Use simple, interpretable models (e.g., small decision trees) to approximate complex black box models.
Example-Based Techniques	Use representative examples or prototypes to explain model behavior.
Counterfactual Explanations	Suggest minimal changes to input features that would lead to a different model prediction.
Rule-Based Methods	Provide interpretable rules describing how the model makes decisions.
Visualization Techniques	Highlight important regions or features using heatmaps, saliency maps, or attention plots.

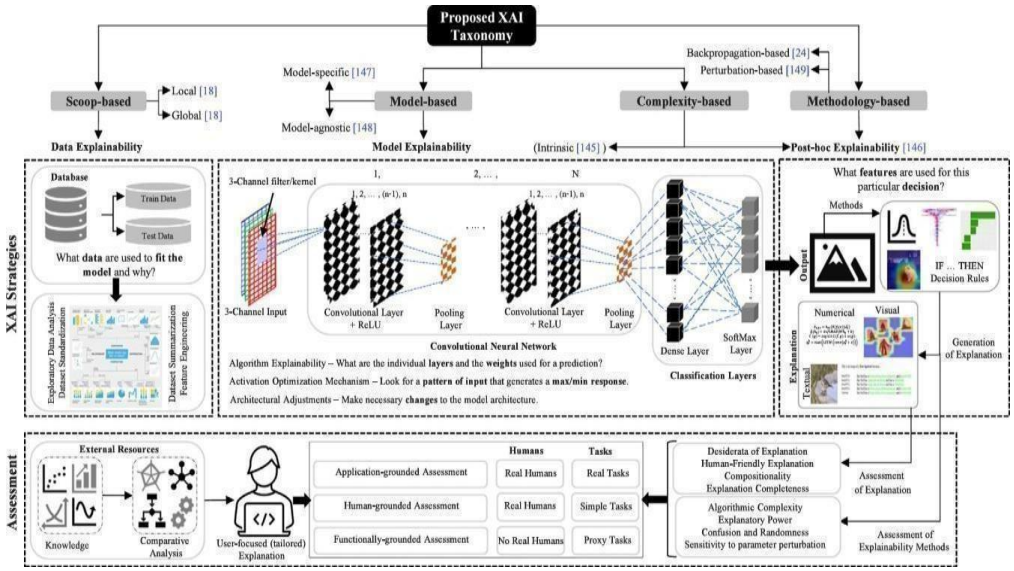


Fig 2. Four-axis XAI methodology

- **Transparency:** In order to comprehend how decisions are made, AI systems should generate outputs that are transparent, intelligible, and interpretable by users. Achieving transparency helps stakeholders understand how the model behaves in high stakes

situations, builds trust, and supports proper oversight.

- **Human Oversight:** The explanations should enable operators to understand and evaluate automated decisions, and also to intervene or override them when necessary. It ensures that final actions are informed and well reasoned.
- **Dataset Documentation:** XAI helps us check the data. It shows if the data has mistaken, too few examples, or hidden unfair things that can change the model's answer.
- **Fairness and Bias Auditing:** Attribution and feature importance methods help to identify discriminatory patterns or reliance on sensitive attributes and support fairness evaluation in AI systems.
- **Traceability and Accountability:** Rule based models, counterfactual explanations, and surrogate models support the creation of auditable decision trails.

Works such as Hummel et al. [17] demonstrate how XAI supports regulatory compliance in domains like clinical decision support systems by offering interpretable reasoning for high risk predictions.

5 Evaluation Metrics for XAI

Evaluating explanations in XAI is not straightforward, since an explanation must both reflect the model's true behaviour and be understandable to a normal user. The following are some widely used criteria for assessing the quality of explanations:

- **Fidelity:** Checks whether the explanation accurately reflects the model's behaviour. Example: If a LIME surrogate model produces outputs similar to the original model, the explanation has high fidelity.
- **Stability:** Ensures that similar inputs lead to similar explanations. Example: Two nearly identical medical cases should not produce very different SHAP values.
- **Simplicity:** Explanations should be concise and easy to follow. Example: A short rule based explanation is easier to understand than a long, complex one.
- **Completeness:** Determines whether all important factors behind the prediction are included. Example: Integrated Gradients distributes importance across all features, offering a complete view.
- **Interpretability:** The explanation must be something a user can easily make sense of.
- **Example:** A heatmap is often more interpretable than a large numeric table.
- **Usefulness:** Measures whether the explanation helps users understand outputs or detect errors in the model. Example: An explanation that highlights feature errors can guide model debugging.

These criteria sometimes reinforce one another, and sometimes conflict. For instance, an explanation that is very simple may lose fidelity. Therefore, the choice of metrics depends on the application context and the types of users who will rely on the explanations.

6 Software Toolkit for Explainable AI

A number of open source toolkits support practical implementation of XAI. Table 2 summarizes the most widely used frameworks, their core capabilities, supported models, and known limitations.
These toolkits allow for practical adoption of interpretability methods across domains and support debugging, auditing, fairness assessment, and building trust in real ML pipelines.

Table 2: Comparison of Common XAI Software Toolkits

Toolkit	Key Capabilities	Supported Models	Limitations
SHAP	Shapley value–based feature attribution; global and local interpretability; optimized TreeSHAP for tree models.	Tree models, deep neural networks, linear models, and general black-box predictors.	Computationally expensive on large datasets; explanations can be slow for deep models.
LIME	Local surrogate explanations; linear interpretable approximations around specific predictions; simple visualizations.	Model-agnostic; works with any classifier or regressor.	Explanations can be unstable; sensitive to sampling randomness; unsuitable for global interpretation.
AIX360	Interpretable models (e.g., rule lists); attribution tools; fairness metrics; robustness diagnostics.	Neural networks, tree models, linear models; supports intrinsic and post-hoc interpretability.	Framework complexity; requires familiarity with IBM libraries; limited visualization support.
Captum	Deep learning–oriented attribution methods: Integrated Gradients, DeepLIFT, Grad-CAM, saliency maps.	PyTorch-based neural networks.	Restricted to PyTorch; unsuitable for tree models, ensembles, or non-neural architectures.

InterpretML	Glass-box models (e.g., Explainable Boosting Machine); feature importance; global and local dashboards.	Tree-based models, linear models, EBM; black-box models via LIME/SHAP integration.	Dashboard-focused; limited support for advanced deep learning; fewer causal/fairness tools.

7 XAI: Challenges and Limitations

Even though XAI is getting better, it still has many problems, so we can't use it easily in real work.

- **Accuracy–interpretability trade off:** Simple and transparent models are easier to understand, but they often do not achieve the high accuracy of deep neural networks or ensemble methods, especially on complex tasks.
- **Lack of standardization:** Different studies use different XAI techniques and evaluation methods, making it difficult to compare results or agree on what qualifies as a high quality explanation.
- **Inconsistent behavior across models**: Multiple explainability techniques produce different results depending on the underlying model architecture. Such an explanation that appears reliable for a tree based model might fail to provide meaningful insights for a neural network which highlights a lack of cross model consistency.
- **Human centered constraints**: Policymakers, loan officers, and physicians could not understand explanations designed for developers. Users might misunderstand or have too much faith in explanations.
- **Problems with robustness and stability**: In real world multiple applications there are where choices must be consistent and even slight modifications to the input data can result in dramatically different explanations regardless of the model's confidence.
- **Scalability issues**: Millions of parameters in deep models necessitate significant processing power to produce explanations, which slows deployment and restricts viability.
- **Risks to privacy and security**: Certain explanation strategies could expose models to adversarial usage, reveal trends from the training data, or leak sensitive information.
- **Uncertainty in regulations**: Despite the fact that policies or laws such as the EU AI Act stress the importance of transparency but still debate regarding the extent and particularity of the justifications needed to achieve full compliance.

These are some of the drawbacks that emphasize the continuous need for more robust standards, trustworthy evaluation processes, and improved technical explanations that better reflect users' actual demands.

8 Conclusion

With the high use of AI systems in critical sectors such as public services, healthcare, and finance, the need for transparent and reliable explanations has shifted from a luxury to a necessity. Explainable AI (XAI) aims to provide insights into complex "black box" models which can also refer to opaque systems in which many existing explanation methods remain difficult for users to interpret, and often lack consistency and trustworthiness. Some of following measures should be fulfilled in the branch of AI explainability:

- **Sophisticated algorithms** which are capable of generating precise, interpretable, and actionable explanations for complex AI models.
- **Clear evaluation standards and metrics** for ensuring that explanations are meaningful, reliable, and comparable across applications.
- **Practical guidelines and domain specific tools**, for unique requirements of fields such as medicine, finance, and law—for such as enabling a doctor to understand which symptoms influenced a diagnosis, or a loan officer to know the reasons for approving or denying an application of users.

The explanations must be understandable, pertinent, and loyal to model thinking instead of complexity of AI models. XAI development should also respect fundamental society values like justice, privacy, safety, and transparency to fulfill and guarantee reliability.

Following things should be considered in future study such as:

- **Interactive and causal explanations** which let people to investigate "what-if" scenarios and comprehend the effects of changing variables.
- **Domain adaptive XAI approaches** which are specifically designed for high stakes applications. Human centric evaluation of explanations is supported by robust, consistent evaluation processes.
- **Integration with regulatory frameworks** which guarantee compliance and also preserve usefulness.

Providing explanations that faithfully represent model reasoning and successfully assist well informed decision making is the ultimate objective of XAI. By addressing current limitations and pursuing these future directions, XAI can make AI systems fairer, more transparent, and accountable, enabling their responsible deployment in real world application.

References

[1] Ali, S., Abuhmed, T., ElSappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. 99, 101805 (2023)

[2] Ortigossa, E.S., Goncalves, T., Nonato, L.G.: Explainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. IEEE Access (2023)

[3] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., Stumpf, S.: Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. arXiv preprint arXiv:2310.19775 (2023)
[4] Pawar, N.A., Mukhmale, S.G., Mule, M., Madel, S.A.: Explainable Artificial Intelligence (XAI). International Journal of Engineering Research (2023)
[5] Srihari, S.N.: Explainable Artificial Intelligence: An Overview. Department of Computer Science and Engineering, University at Buffalo (2020)
[6] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable Artificial Intelligence: A Brief Survey on History, Research Areas, Approaches and Challenges. Lecture Notes in Computer Science. Springer (2019)
[7] Mathew, D.E., Ebem, D.U., Ikegwu, A.C., Ukeoma, P.E., Dibiaezue, N.F.: Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Humans. Journal of Artificial Intelligence and Data Science (2025)
[8] Rane, N.L., Choudhary, S.P., Rane, J.: Explainable Artificial Intelligence (XAI) in Healthcare: Interpretable Models for Clinical Decision Support. International Journal of Healthcare Informatics (2024)
[9] Černevičienė, J., Kabašinskas, A.: Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review. Financial Innovation. 10(3), pp. 1-28 (2024)
[10] Kohan, A., Zahedi, A., Alizadehsani, R., Tan, R.S., Acharya, U.R.: Application of Explainable Artificial Intelligence (XAI) Techniques in Patients with Intracranial Hemorrhage: A Systematic Review. Expert Systems with Applications (2025)
[11] Ribeiro, M.T., Singh, S., Guestrin, C.: 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 (KDD), San Francisco, CA, USA, pp. 1135-1144 (2016)
[12] Rodis, N., Sardianos, C., Radoglou-Grammatikis, P., Sarigiannidis, P., Varlamis, I., Papadopoulos, G.T.: Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions. arXiv preprint arXiv:2306.05731v2 (2024)
[13] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for Large Language Models: A Survey. arXiv preprint arXiv:2309.01029v3 (2023)
[14] Okonkwo, R., Folorunso, A., Ogundipe, F., Tettey, C.Y.: Explainable Artificial Intelligence through Human–AI Collaborative Frameworks: Quantifying Trust and Interpretability in High Stakes Decisions. Computer Science & IT Research Journal. 6(5), pp. 333-354 (2025)
[15] Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.R., Wolf, L.: XAI for Transformers: Better Explanations through Conservative Propagation. arXiv preprint arXiv:2202.07304v2 (2022)
[16] Tu, N.V., Long, P.N.H., Viet, V.H.: xAI-CV: An Overview of Explainable Artificial Intelligence in Computer Vision. arXiv preprint arXiv:2509.18913v1 (2025)
[17] Hummel, A., Burden, H., Stenberg, S., Steghöfer, J.P., Kühl, N.: The EU AI Act, Stakeholder Needs, and Explainable AI: Aligning Regulatory Compliance in a Clinical Decision Support System. XITASO GmbH & RISE Research Institutes of Sweden, University of Bayreuth (2025)