Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explainable Artificial Intelligence (XAI): Promoting the transparency and trust in machine learning models.

Archy Biswas[1†], Utsha Sarker[1*†], Lalit Vaishnav[1], Ikram Ali[1], Ruksana[1], Navjot Singh Talwandi[1]
[1]Department of CSE, Apex Institute of Technology, Chandigarh University, Gharuan, Mohali, 140413, Punjab, India.
*Corresponding author(s). E-mail(s): archyz2021@gmail.com ; utsha.sarker00775@gmail.com and navjotsingh49900@gmail.com;
Contributing authors: *vlalith7036@gmail.com* ; ikram425ali@gmial.com ; ruku78601@gmail.com ;

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

The swift proliferation of machine learning and deep learning models into the domain of sensitive areas like healthcare, finance, and the law has created a great worry about their lack of transparency and interpretability. These models tend to run as black box, thus restricting their usage in high stakes situations as the users might not understand, trust or even validate their decisions [10], [13]. Explainable artificial intelligence (XAI) has become a promising paradigm to deal with these two issues by making models explain themselves in human understandable terms and not engaging in predictive performance to a considerable degree [33].

This paper provides an in-depth and organized summary of XAI along with its major concepts in relation to interpretability, transparency, trust, and accountability. It also divides XAI techniques into intrinsically interpretable (ante-hoc) models and post-hoc explanation methods and differentiates between model-specific and model-agnostic [13], [39]. The paper also looks at the association between various kinds of explanations and user confidence in various fields of applications.

It suggests a new conceptual framework that demonstrates the connection between technical attributes of explanations, including fidelity, stability, and completeness, and human-related aspects, including trust, understanding, and fairness. The main contributions made by this work are: a taxonomy of XAI methods are complete aligned with model architectures and data modalities, a more integrated framework with explanation types, their target users, and measures of evaluation, provides deeper insight into the open issues of standardization of evaluation, faithfulness of their explanations and regulatory adherence [38].

In general, the proposed research should inform researchers and practitioners to choose, design, and assess XAI methods to improve the transparency, reliability, and trust in the current AI systems.

Keywords: Explainable Artificial Intelligence (XAI), Interpretability, Transparency, Trust in AI, Post-hoc Explanations, Interpretable Models, Model-agnostic XAI, Human-centric AI.

1 INTRODUCTION

Blistering development of machine learning (ML) and deep learning (DL) technologies has been very influential in changing the modern artificial intelligence (AI) systems, which have become quite accurate predictors in various fields including healthcare, finance, cybersecurity, and public policy. Even state-of-the-art models, especially deep neural networks and ensemble algorithms, although high-performing, are opaque, black-box systems, meaning the way the information is handled to form an internal decision can often not be easily understood by humans [10], [13]. This obscurity presents a crucial challenge to their extensive use in high stakes contexts, where the explanations that underlie predictions count equally to the predictions themselves.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The clinicians need to be able to get a definitive explanation behind the AI-assisted diagnosis, especially in medical fields like healthcare to maintain the safety of patients and uphold ethics. Equally, in the world of finance, regulatory authorities insist on transparency of the automated decision-making systems to avoid bias and enhance fairness. Making algorithmic decision biases justifiable in law and policy-making processes is critical to accountability and trust in the community. The resulting interpretation problem of complex ML models compromises its reliability, restricts its adoption in users, and also makes investigating and diagnosing errors difficult [21], [33].

The issue of openness is also exacerbated with changing regulatory frameworks. As an example, legislation like the General Data Protection Regulation (GDPR) highlights the right to explanation, which obliges organizations to give purposeful details about automated choices that influence people. This has heightened the pressure on the development of AI systems which are accountable, interpretable and accurate. Lack of these abilities may result in ethical issues, decreased trust, and even legal consequences, especially when AI systems are involved in making a decisive conclusion [27], [38].

Explainable Artificial Intelligence (XAI) has become a major area of research in order to overcome these issues and encourage the transparency and readability of artificial intelligence systems. XAI involves a collection of methods and approaches that allow people to comprehend, believe in, and successfully oversee AI-backed choices. In a general sense, there are two broad classes of XAI methods:

(i) intrinsically interpretable (ante-hoc) models, including decision trees, linear models, which are modeled to be transparent, and

(ii) post-hoc explanation methods, which seek to provide explanations of the behavior of complex black-box models following training. Post-hoc methods may be further subdivided into model specific methods (e.g. gradient based explanations of neural networks), and model agnostic methods (e.g. LIME and SHAP) which may be applied to a wide variety of models [13], [39].

Most of the current surveys are predominantly discussing technical taxonomies in the world, algorithmic advances, or specific application domains, although a considerable literature has been developed on XAI in different aspects. Although these contributions are worthy, they tend to dismiss the importance of the characteristic of explanation and human oriented variables such as trust, usability and properly made decision. Specifically, the research that systematically relates the type and quality of explanations to the effects they have on various stakeholders such as developers, domain experts, end-users, and regulatory authorities is not much [12], [33]. This gap underlines the necessity of its more comprehensive and interdisciplinary approach uniting both technical and human-focused aspects of explainability.

To counteract these drawbacks, this paper will set out to give a thorough and systematic examination of XAI and lay significant stress on transparency and trust. To start with, it presents the conceptually clear difference between such central concepts as interpretability, transparency, and trust in the relation to the ML systems. Whereas the interpretability is a measure of how much a human can know about the inner-workings of a model, transparency is concerned with the openness and availability of the structure and behaviour of a model. Trust however is a human aspect that is subjective based on quality, clarity and reliability of explanations [10], [38].

Second, the work offers a more detailed taxonomy of XAI techniques based on model types or data modalities (e.g., tabular, image, text) and explanation targets (e.g., global or local explanations). This taxonomy helps to gain a dynamic perspective on the application of different methods in diverse situations and elucidate their advantages and drawbacks.

Third, to close the divide between technical properties of explanations, including fidelity, stability, and completeness, and human-related outcomes such as trust, understanding and the perceived fairness, a novel conceptual framework is suggested based on trust. This framework is a systematic way of studying XAI methods regarding not only their technical performance, but their role in practical, real-life contexts of decision-making.

Lastly, the paper summarizes several outstanding issues and future research fronts in XAI. These encompass the necessity of standard measures of evaluation, faithfulness of explanations, the presence of biases in explanations, as well as attainability of compliance with the new regulatory requirements. Interdisciplinary collaboration as critical

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

to future development of the f ield to highly trustworthy AI systems is also discussed [38], [39].

In general, the research should assist scholars and practitioners to choose, design, and evaluate XAI methods to increase transparency, accountability, trust in AI systems and achieve responsible implementation of its use in high-stakes applications.

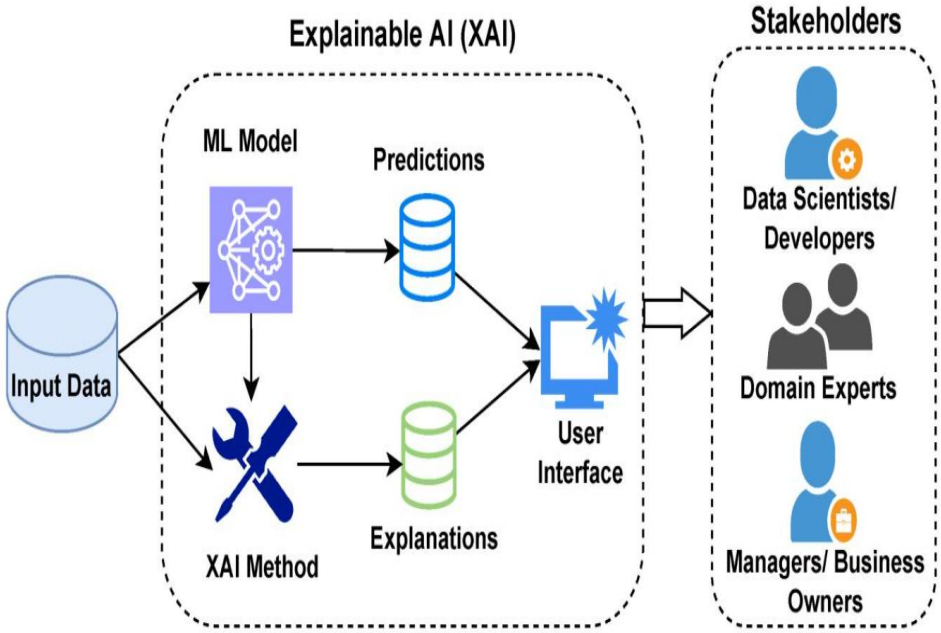

Fig. 1. XAI within the AI life cycle. The classical machine learning pipeline (data (input) model (output) prediction) is expanded with an explanation layer giving interpretable information to various stakeholders, such as developers, domain experts, end-users, and regulators that allows transparency, trust, and accountability [33], [38].

2 PRINCIPLES: OPENNESS, EXPLAINABILITY AND CONFIDENCE

2.1 Definitions and Rich Employees

The growing number of applications of machine learning (ML) systems in high stakes settings has driven a need to have a clear perspective on the core concepts of interpretability, explainability, transparency and trust. These ideas have the theoretical foundation of Explainable Artificial Intelligence (XAI) and are key to ensuring that the systems are not only right, but also responsible and people-centered.

Interpretability is the extent of the human cognisability of the inner workings of a model or the kind of process that the model follows in making a decision. Linear regression, decision trees and rule-based model are the few interpretable models that can provide an easy user trace of how input features can impact outputs. Conversely, deep neural networks among other complex models, are usually not inherently readable [13], [33] because they are highly dimensional and non-linear.

Explainability is a concept which is more general in contrast to interpretability though they share similarity. It refers to the capability of a system to give human understandable explanations to its predictions, in any case the model behind it may or may not be interpretable. This can be accomplished by post-hoc methods that come up with explanations not intrinsic to the model as seen in feature attribution techniques, surrogate models as well as counterfactual explanations [39]. Therefore, a model can be non-interpretable, but can be explained with the help of the right methodology.

Transparency involves openness and availability of information related to the model such as the model architecture, its training information, feature representations, and decision logic. Transparent systems allow the stakeholders to inspect, audit, and validate behaviors of the model thus facilitating reproducibility and accountability. The process of transparency is quite critical when it comes to regulated settings where compliance and ethics are of paramount importance [27], [38].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Trust is a human-oriented concept which indicates the desire of the user to trust an AI system. Various factors affect trust, such as perceived competence (accuracy and, performance), reliability (remains consistent over the inputs), fairness (is free of bias), and clarity (can be understood by the user). Notably, trust does not constitute entirely based on the technical performance, but it is influenced by how convincingly the system reasons to various stakeholders [10], [12]. XAI in this case is the networking between algorithmic complexity and human intelligibility, as well as builds trust via significant and trustworthy explanations.

These are four concepts which are related, yet different. Interpretability and explainability are more concerned with the how of comprehending model choices, whereas transparency refers to the information availability, and trust is the views and connections of the user to the system. A thorough XAI framework should take into account each of these dimensions to have effective and responsible AI deployment.

2.2 Types of Explanations

XAI techniques produce types of explanation based on the intended audience, area of application, and the complexity of the model. These explanations can be sorted out on various dimensions.

- **Explanations Local vs. Global:**

Local explanations also seek to explain individual predictions in terms of what contributions as a feature made to a particular output. Such methods as SHAP and LIME are popular to do it. Global explanations, on the other hand, give information about the entire behavior of a model including the rankings of features in importance or decision limits. Local explanations are helpful to the analysis of cases on a case-by-case basis, but global ones are needed to validate a model and make decisions at policy-level [13], [39].

- **Level-based, Concept-level and Example-based Explanations:**

The feature-level explanations concentrate on the impact of single input variables to a prediction. These have widespread use in tabular data applications and common usage in feature attribution various techniques. Explanations of concepts, however, are more abstract, and combine features into concepts comprehensible to humans (e.g., the shape of the tumor, in medical imaging). Explanations in form of examples give representative examples, prototypes, or counterfactuals of how the input varies with output. They especially work well in making decisions and user intuition better [33], [38].

- **Explanations vs. Explanations:**

Technical vs. Human-Centric Technical descriptions have been made to serve developers and researchers and are usually mathematical formulations, gradients or internal model parameters. These are accurate explanations but might not be comprehended easily by the non-experts. Human-centric explanations contrastingly, are designed to satisfy the cognitive and contextual requirements of the user, and offer simplified, narrative or visual explanations, consistent with domain knowledge. These explanations can be important to end-users and decision-makers that need to receive actionable information but not technical details [12], [33].

Multifaceted nature of types of explanation exposes the need to have context sensitive XAI systems that can generalize explanations to various stakeholders. An example is a data scientist who might need more impressive scores on feature attribution as opposed to a clinician who might need a brief summary of important medical aspects of influencing a diagnosis. Equally, regulators might require explanations to determine fairness and compliance across the globe.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

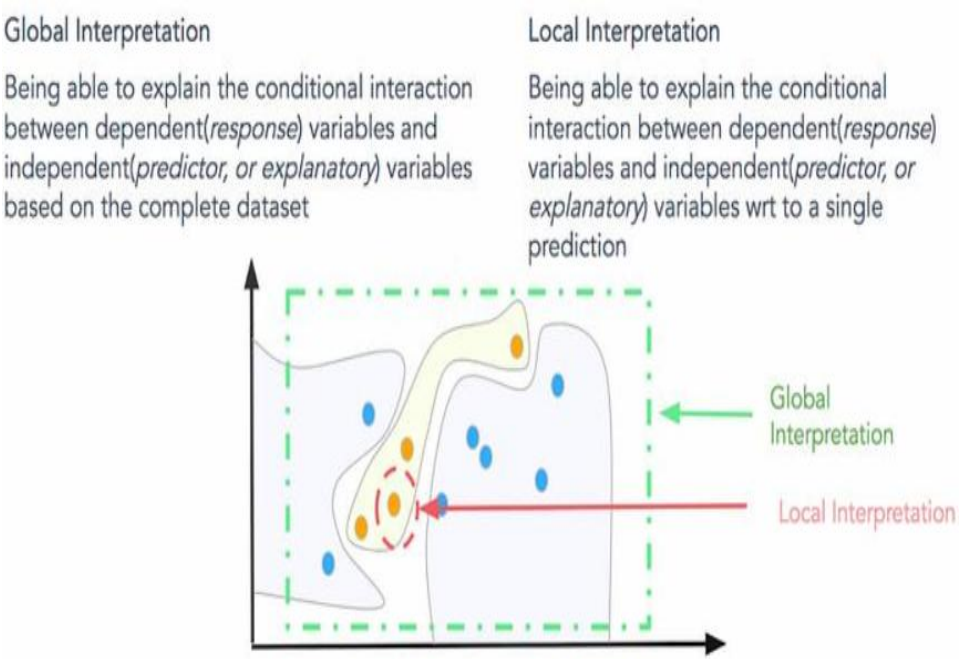

Fig. 2. Explanation space in XAI. Two-dimensional scheme that depicts the range between local to global explanations and feature-level to concept-level explanations to illustrate the range of the type of explanations and its applicability [13], [39].

Explanation Type	Typical Audience	Example Methods	Typical Use-Cases
Local, Feature-level	Developers, Data Scientists	LIME, SHAP	Debugging individual predictions, feature attribution
Global, Feature-level	Regulators, Analysts	Feature importance, PDP	Model validation, policy compliance
Local, Concept-level	Domain Experts	TCAV	Interpreting domain-specific patterns
Global, Concept-level	Researchers, Regulators	Rule-based summaries	Understanding behavior at abstraction
Example-based (Local)	End-users	Counterfactuals	Decision support
Example-based (Global)	Analysts	Clustering	Pattern discovery
Technical	Developers	Gradients, saliency maps	Debugging, optimization
Human-centric	End-users	Visualizations, narratives	Trust-building

TABLE I. The categories of explanation found in XAI such as local/global and feature/concept/example based explanations are summarized and mapped into common stakeholders, methodology, and applications, underscoring the design-user gap in explanation design and user requirements [13], [33], [39].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3 TAXONOMY XAI METHODS

The increasing need to have transparency in machine learning (ML) systems has resulted in a host of Explainable Artificial Intelligence (XAI) methods. All these approaches can be generally divided into three large families

(i) intrinsically interpretable (ante-hoc) models,

(ii) post-hoc model-specific model explanation methods, and

(iii) post-hoc model-agnostic models. This taxonomy offers a systematic insight into the manner in which explainability is possible at various steps in the modeling pipeline and indeed different model types and models in different data modalities.

3.1 Intrinsically Interpretable Models (Ante-hok)

Transparent by design, intrinsically interpretable models are created with the aim of being understood by users about how input features are used to make predictions. The models tend to be favored by the applications where interpretability is a major concern and this may include health care diagnosis, financial decisions and regulatory system [13], [33].

Some common intrinsically interpretable models are decision trees, rule-based systems, linear and logistic regression models, sparse linear models and generalized additive models (GAMs). Decision trees offer a hierarchical format of decisions which can be readily displayed and comprehended whereas rule-based models illustrate decisions as human-conceptualized, readable rules of condition (if) and consequence (then). Linear models, especially when they are sparsified, provide straightforward insights into the importance of the features, in terms of their coefficients.

More sophisticated interpretable architectures are also being suggested in recent years. They include prototype based models that group the inputs according to their similarity with representative examples; concept bottleneck based models that incorporate explicitly human interpretable concepts in the prediction pipeline. These are meant to overcome the awkward trade-offs existing between high performance and interpretability, by integrating semantic reasoning into the model itself [39].

The high transparency and easy explainability of intrinsically interpretable models is the main strength since it avoids extra levels of explanation. They specifically work well with structured/tabular data, and there are fairly simple relationships between variables. In addition, such models facilitate simpler debugging and fairness inspection as well as regulatory compliance.

Nevertheless, these models have significant shortcomings too. They are less complex and therefore may not be able to capture complex, non-linear relationships in high-dimensional data like images, audio and natural language. Consequently, they might not be as effective as deep learning model in tasks that demand intricate feature extraction and representation learning [10], [33]. This interpretability-performance trade-off is still one of the key issues of XAI studies.

3.2 Post-hoc Model-Specific Methods

Post-hoc model-specific model-specific methods explain the model behavior after training based on the knowledge of the internal structure of the model. These methods are especially applied to deep learning models, in which the intrinsic interpretability can be unavailable.

Various popular explanation methods have been created in the context of neural networks. Gradient-based approaches (like Integrated Gradients and Grad-CAM) and saliency maps indicate the strength of input features (e.g., pixels in an image) through analysis of gradients with respect to the output. The techniques offer visual explanations that are particularly helpful in machine vision problems so that a user can tell which parts of an image were used to make a prediction [11], [13].

Another kind of interpretability is the attention mechanisms in the case of natural language processors (NLP). The weights of attention can be visualized to show which words or tokens the model is paying attention to in order to make a prediction. Even though the explanations, which use attention as an intuitive explanatory variable, are intuitive, the fact that they are reliable as faithful explanations is a topic of continuing controversy [33].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The other notable method is the Layer-wise Relevance Propagation (LRP) method, which breaks down the prediction of a neural network, propagating scores of relevance back down through the layers. LRP has a more formal description than simple gradient-based algorithms and is typically applied in situations where it is crucial to be able to interpret a model like in medical imaging [13].

On tree-based ensemble models (e.g., Random Forests, Gradient Boosting Machines), feature importance measures or partial dependence plots (PDP), and surrogate decision trees are model-specific models that provide model explanations. Importance of features measures the value added by each feature to the model prediction whereas PDPs depict the interaction between a particular feature and the outcome prediction. Surrogate models are models that can simulate the behavior of complex ensembles using interpretable, simple, models [13], [39].

The primary benefit of model-specific methods is that they are highly fidelitous in that they make use of the inner workings of the model to produce an explanation. Nevertheless, they are restricted to certain classes of models, which decreases their versatility. Also, certain algorithms can derive the explanations that could not be easily interpreted by non-experts, especially when most of the data is high-dimensional.

3.3 Post-hoc Model-Agnostic Methods

The objective of post-hoc model-agnostic techniques is to give an explanation, which does not depend on the internal property of the model and is applicable to any black-box system. These are also very adaptable techniques with extensive application in practice, particularly when a mixture of different types of models are utilized.

Perhaps one of the most noticeable model-agnostic methods is the LIME (Local Interpretable Model-agnostic Explainations) which attempts to approximate the actions of a complex model in a localized and simpler form by the use of an interpretable surrogate model. On the same note, SHAP (SHapley Additive exPlanations) is their counterpart, and is also founded on cooperative game theory, with its main distinction that it uses each feature to give a contribution value as to its significance in a prediction. SHAP offers local and explanations on a global scale and it has been highly adopted because it has excellent theoretical background [13], [39].

Another significant type would be a set of surrogate models, in which a interpretable model (e.g., a decision tree model) is trained to approximate the behaviour of a black-box model. Although practical in obtaining a global understanding, surrogate models can be affected by the approximation errors and hence, this can result into a potential discrepancy between the surrogate and original model.

Counterfactual explanations may be useful in offering insights by determining the minimal differences to input features, which would change the model prediction. Such explanations are especially useful in the context of decision support as they can provide recommendations to be acted upon (e.g., increase income by X to receive loan approval). Model behavior can be explained using examples, such as prototypes and criticisms, to demonstrate to the user how the model works by showing examples [33], [38]. The main advantage of model-agnostic approaches is that they can be easily universalized and can operate on a wide range of models and data. Yet, those approaches frequently base on approximations, a fact that may cast doubt on the reliability and accuracy of explanations. Also, they can add to computational overhead especially with large models and datasets.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

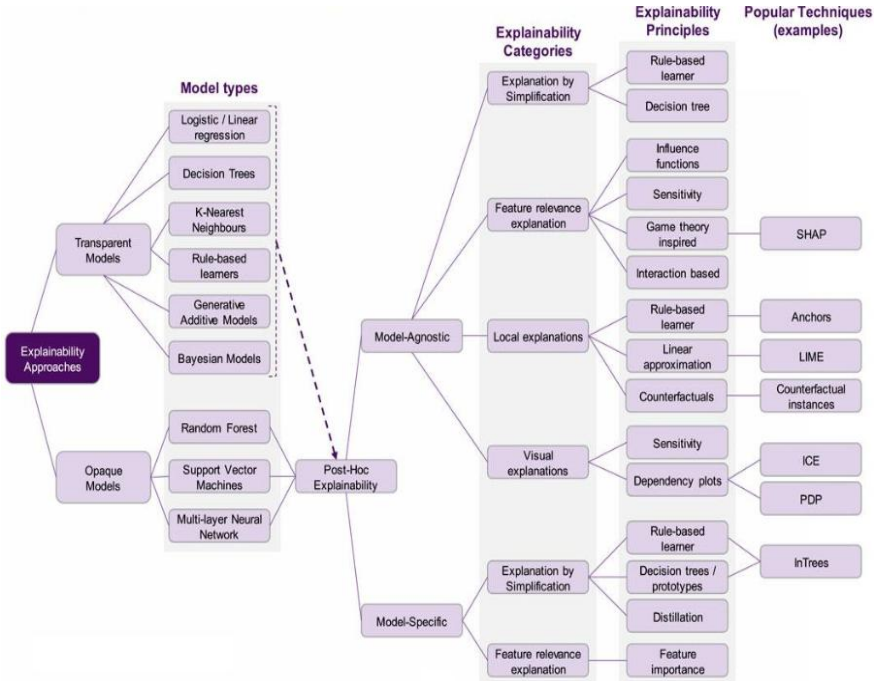

Fig. 3. XAI method families. Surgery of the three most significant categories of explainability approaches, namely intrinsically interpretable models, model-specific post-hoc model-specific techniques and model-agnostic post-hoc model-agnostic ones, characterizing how they formulate explanations to various types of AI models [13], [39]

Category	Example Methods	Supported Models	Supported Data	Pros	Cons
Intrinsic	Decision Trees, GAMs	Interpretable	Tabular	High transparency	Lower performance
Post-hoc Model-Specific	Grad-CAM, IG, LRP	NN, Ensembles	All	High fidelity	Complex
Post-hoc Model-Agnostic	LIME, SHAP	Any	All	Flexible	Approximation errors

TABLE II. XAI approaches are classified into intrinsically explainable models, post-hoc model-specific methods, and post-hoc model-agnostic approaches, and supported type of model, data modalities, benefits, and drawbacks [13], [39].

4 XAI OF TRANSPARENCY AND TRUST PRACTICE

4.1 Application Domains

The real effect of Explainable Artificial Intelligence (XAI) can be best seen in high stakes areas where transparency, accountability, and trust are crucial when it comes to the decision-making process. XAI practices would inherently be generalized, but their performance will depend on the domain demands, the expectations of the concerned parties, and the regulations.

- **Healthcare:**

The AI systems find applications in diagnosis, prognosis, treatment plan, and clinical decisions in healthcare. The use of black-boxes in clinical environments is however restricted because of the necessity to possess an

interpretability and accountability. To guarantee patient safety and ethical standards, clinicians should be aware of why a model behaves in a given manner. Saliency maps (i.e., Grad-CAM) can be used as an XAI in medical imaging, feature attribution algorithms (i.e., SHAP) can be used to identify these areas of interest and risk factors in clinical data analysis [1], [20].

To illustrate, an example of XAI in radiology is showing areas of tumors that affect a diagnosis, whereas in predictive analytics, XAI can show important variables through variables like age, blood pressure, and genetic markers. These explanations not only increase the confidence of the clinicians, but also help in the process of validation and contributing to the decision making [21], [22].

With these positive features, however, there are still issues such as the provision of clinical validity and reliability of explanation and harmonizing it with the knowledge of the domain. The explorations should be brief, medically significant and consistent across cases to be of use, practically.

- **Finance:**

XAI in the financial sector is critical in making sure that there is fairness, transparency, and that the application is free of regulatory abuses. Credit scoring, loan approval, fraud detection, and risk assessment are some of the most common credit score, loan approvals, and risk assessment tasks performed by AI models, as they are life-changing to individuals and organizations. Turning to automated decisions can be a legal requirement by regulatory authorities who may ask institutions to explain happenings in automated decisions [40].

The SHAP technique is a typical feature attribution technique that is employed to describe credit decisions and quantify the impact of variables, such as income, credit history and debt ratio. Counterfactual explanations are helpful in the given area, in particular, especially useful since they offer actionable information (e.g., reduce debt by X to qualify better to take a loan). Also, to ensure model fairness and to identify biases among groups of people, global explanation methods could assist the auditors in their practice [16], [40].

Nevertheless, financial explanations should be as accurate as they are understood with a balance between the technical accuracy and interpretation, so that they can be understood by specialists (e.g. credit analysts), and non-experts (e.g. customers). The domain of transparency is also strongly related to the concept of trust because there is a risk of reputation and lawsuits in case of opaque decisions.

- **Law Public Sector:**

Application of AI in people sector and the legal frameworks has brought up serious issues relating to transparency, fairness and accountability. Their uses are in risk assessment in criminal justice, resource allocation, distribution of welfare, and decision-making on policies. Decisions used in these situations can have a significant societal impact and explainability is frequently the only way to guarantee a certain level of trust by the population and guarantee ethical governance [27], [38].

The application of XAI techniques facilitates clarity of the reasoning that led to the decision and therefore stakeholders like policymakers, legal experts and citizens will be able to comprehend and are allowed to question the results. Rules-based explanation and global models can be used to justify policy decisions on one hand and local explanations can explain the outcome of individual cases, on the other hand.

One of the very important issues in this field is the necessity to guarantee the technical correctness of the explanations as well as to make them socially and ethically acceptable too. They need to be explained with references to issues of bias, discrimination, and fairness especially in cases where AI systems refer to the vulnerable population. Further, regulatory efforts are driving toward greater transparency and supporting the significance of XAI in the use of the technology in the public sector.

- **Cross-domain Observations:**

In all fields, the suicidal nature of XAI is contingent upon the capability of adapting towards the various stakeholders. The detailed technical explanations that developers might need to provide to debug their work might be more suitable to the end-user and decision-maker, who might be more interested in simplified and context-sensitive explanations. This points to the significance of human-centric design to XAI where the explanations must be user-centered and

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

also domain based [12], [33].

4.2 The role of Explanations in affecting Trust

Explainability and trust relationship is a complex and multifaceted relationship. Although transparency is traditionally assumed as a precondition of trust, empirical research indicates that the explanations should satisfy certain qualitative standards to provide users with a certain level of trust towards AI systems.

First, there is contextual relevance which is crucial. Explanations explaining in line with its domain knowledge and decision making circumstances are more likely to be trusted. As an illustration, clinicians are more likely to appreciate an interpretation with a focus on medically significant attributes whereas the financial users would appreciate information regarding risk factors, and the rates of compliance. Explanations that are generic or even overly technical, can be unsuccessful even when accurate [10], [12].

Second, it is imperative that it is concise and clear. Maintaining simplicity in the manner the explanations are given to the users is likely to attract them rather than complexities. Explanations that are too detailed may end up confusing the user and decrease his/her capacity to interpret the results. Especially when it comes to non-expert users, this is crucial because they can have no clue in interpreting complex output when lacking the technical background [33].

Third, a factor that plays with trust is consistency and stability. There should be consistency in explanations when dealing with similar inputs and these explanations should not significantly change with small changes in data. It is the explanations that fail to be consistent that can be used to weaken confidence in the model as such predictions may be accurate [38].

Fourth, technical properties that are important are fidelity and faithfulness. The way the model should be explained should reflect the actual behavior it should exude instead of giving an approximation that is not very realistic. Explanation with low-fidelity may give an illusion of trust resulting in overtrust on the system [39]. Notably, studies have shown that being transparent through the use of algorithms is not sufficient to achieve trust. A transparent model does not necessarily earn the trust of users; instead, it is systems, which give understandable, relevant and actionable explanations that earn user trust. This highlights the importance of human-based assessment measures having surpassed technical performance to user perception and quality of decisions [12], [33].

Moreover, the fairness and ethical aspects have influence on trust. Explanations which bring to light biasness or any tendency of discrimination might lessen trust whereas fair and responsible explanations may boost trust. When in regulated areas, action in accordance with ethical guidelines and laws enhances user trust even more [27], [40].

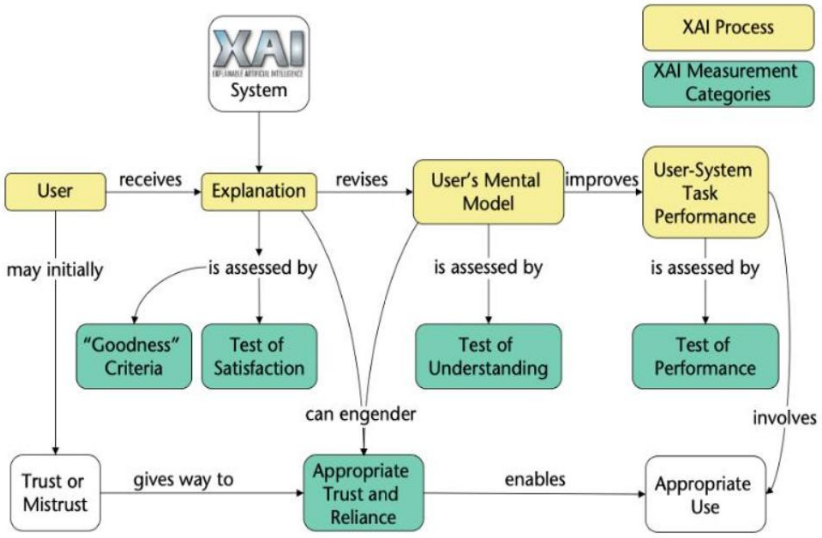

Fig. 4. Technical explanations to user confidence. Pipeline of how XAI techniques generate explanations with fidelity, clarity and relevance, affecting user cognition and ultimately affecting trust and AI systems adoption [10], [12].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Domain	Stakeholder	Trust Requirements	Methods	Constraints
Healthcare	Clinicians	Accuracy, reliability	SHAP, saliency	Ethics
Finance	Auditors	Fairness	SHAP, CF	Regulation
Public Sector	Citizens	Transparency	Rules	Law
Cross-domain	Users	Clarity	Visualizations	Ethics

TABLE III. Comparison of XAI requirements by application domain, stakeholders, trust requirements, the appropriate description techniques and regulatory or ethical issues [21] and [40].

5 ANALYSIS OF EXPLANATIONS: TECHNOLOGICAL AND SOCIAL HUMAN VIEWS

The usefulness of Explainable Artificial Intelligence (XAI) strategies is not just in their capability to produce explanations but it is also assessed based on how the explanations are judged. An evaluation framework needs to be comprehensive in looking at both the technical aspects of the explanations, and human-focused results since they will collectively measure the usefulness, dependability, and credibility of intelligence systems. Whereas technical metrics can be used to determine how faithful and robust a set of explanations is, human-centric metrics can be used to gauge the extent to which explanations aid in understanding, decision-making and trust [13], [33].

5.1 Technical Metrics

Technical assessment of XAI techniques is aimed at exploring the measures of accuracy and reliability of the explanations in the context of the behavior underlying the model. There are a number of important properties and related measures extremely popular in the literature.

- **Fidelity / Faithfulness:**

Fidelity or faithfulness is a measure of how accurately an explanation is of the actual decision making process of the model. High-fidelity explanations make certain that the features or patterns that are brought to the fore by the explanation actually affect the output of the model. In feature attribution techniques, a measure of fidelity might by the ability to remove/perturb critical features and lense the prediction change that occurs. In case of a drastic change in the prediction, then the explanation is said to be faithful [39]. At low-fidelity, it may be misleading to our users and causes them to develop a sense of trust, which is especially troublesome in critical apps.

- **Stability / Robustness:**

The stability determines the similarity of the explanations to the input data when small changes are made to the input data. Ideally when the same input is taken, similar explanations of the same should be come up with. When the slightest fluctuations in input cause considerably diverse explanations then the approach is regarded as unstable, which subverts the reliability of the user. It is particularly in areas like healthcare, where scenarios featuring uneven explanations might result into uncertainty and lack of confidence [38].

- **Completeness / Coverage:**

Completeness is the degree to which the description of the nature of the overall decision logic of the model was completely covered. Although local explanations are the ones that are dwelled on making individual predictions, they can be unable to measure the way the model behaves on a global basis. Completeness metrics typically assess how the explanation is complete in that all the relevant features and interactions which can affect the prediction are explained. Partial accounts can be given leaving out important points thus creating partial or subjective interpretations [13].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

- **Complexity / Sparsity:**

Complexity is a cognitive load having to comprehend a description, whereas sparsity is a measure of features or elements. Explanations that are more concise and simple are preferred; they are not as difficult to interpret by the users. Nonetheless, overly simplified answers might come at the expense of fidelity which is a major trade-off in the design of XAI. Complexity is usually measured by metrics like the number of features that appear in an explanation or depth of a decision tree [33].

These technical measures give a basis of quantification of different XAI approaches. Nevertheless, they are more interested in algorithmic properties and might not overly reflect the perception of the explanations and how they are used by human users.

5.2 Human-Centered Evaluation

Human-centered evaluation aims at evaluating the effects of the explanations to the users regarding their understanding, trust, usability and performance in making decisions. These tests are not technical in nature, but usually tested by conducting user studies, experiments, and qualitative testing.

The evaluation of trust of the user is one of the main goals of human-centered evaluation. Research has preferred that users tend to entrust AI machines by providing clear, relevant, and aligned explanations related to their knowledge on the domain. But trust does not the most depend on the existence of explanations, the quality and presentation of other explanations is also a crucial aspect [10], [12].

The other consideration is usability which assesses the ease of formulating interactions and comprehension among explanations to the users. The tasks associated with usability studies are often the decision making by the users based on model outputs and explanations. Indicators to evaluate usability are usually those that measure things like the time required to complete a task, accuracy, user satisfaction, etc.

Performance in tasks is also a vital measure of effectiveness of explanations. Explainations should be more useful in making the user make right and informed choices. An example is within the medical diagnosis process, explanations are supposed to assist clinicians to locate the relevant features and enhance accuracy diagnosis. On the same note, in finance, justifications ought to help the analysts to single out anomalies and evaluate risks in a better way [21], [40].

Mental model alignment is based on the effectiveness of the description in assisting the users to develop an accurate perception of model operation. An aligned mental model enables users to make decisions based on how the system is expected to act in the various situations, thus enhancing the trust and decision making. Totally inappropriate mental models on the other hand may result in wrong assumptions and relying too much on the system [33].

It has always been empirically proven that the best way to have explanations is through contextualization, conciseness and individualized to the user. The technical or over detailed explanations can prove too much to users and diminutive explanations, may not communicate important information. As such, it is important to design explanations in such a way that there is a trade off between their richness and their clarity in order to have the best effect on user trust and user performance [12], [33].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

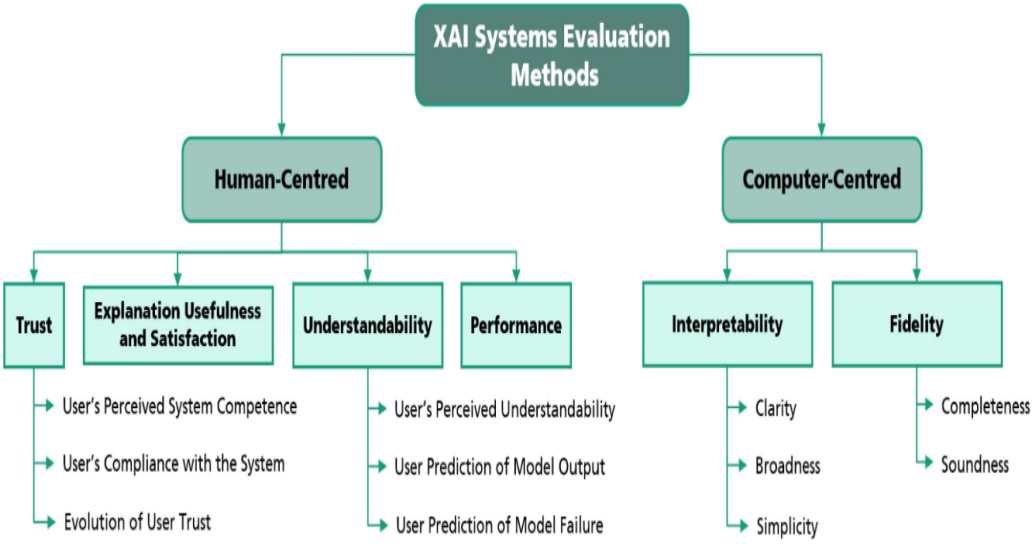

Fig. 5. XAI evaluation Two-layered. The bottom layer is technical assessment measures like f idelity and stability and the top layer is the human-centric measures like trust, understanding and quality of decisions and that integrated assessment is required [33], [38].

Metric	Type	What it Measures	Setup
Fidelity	Technical	Model alignment	Perturbation tests
Stability	Technical	Consistency	Noise injection
Completeness	Technical	Coverage	Comparison
Complexity	Technical	Simplicity	Feature count
Trust	Human	Confidence	Surveys
Usability	Human	Ease	User studies
Task Performance	Human	Decision accuracy	Experiments
Mental Model Alignment	Human	User understanding	Behavioral studies

TABLE IV. Description of the main assessment parameters, such as fidelity, stability, and complexity (technical), and trust, usability, and task performance (human-centric), and their measurement tools [33], [38].

5.3 Evaluation Challenges

Although the area of XAI evaluation has been advanced greatly, there are a number of challenges. Such a conflict of fidelity versus simplicity is one of the most eminent. The explanation that is caused to be high fidelity might be difficult to understand, whereas a simple explanation might not be accurate. The problem of establishing a good balance between these is one of the main research questions [38].

Over-trust is another thereof. Although explanations are supposed to instill trust in a user, it might sometimes cause users to over rate the trustworthiness of the system particularly when the explanations are convincing yet unfaithful. This shows that there is a need to move beyond the method of evaluation that just takes into account trust to the proper method of trust calibration [39].

There are no common measures of evaluation, which also complicates comparisons of XAI approaches. Various researchers adopt various measures and evaluation systems and it proves to be challenging to set benchmarks and generalize results. Also, due to the subjectivity of human-based assessment, there is a degree of variability, with user

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

perceptions varied depending on their background, professional and situational contexts [13], [33].

Lastly, an integrated system of development is needed to incorporate both the technical and human-based approach to evaluation. When assessing explanations we can be interested in both the evaluation of their technical metrics and the evaluation of their practical utility but can also be interested in the assessment of user studies, where we may overlook the properties of importance to the algorithms. To achieve credible XAI systems, it is necessary to commit to a holistic approach featuring the consideration of both of these dimensions.

6 Trust-Centric XAI conceptual framework

6.1 Framework Overview

The growing application of AI to high stakes settings creates the need to have a unified framework that systematically links the nature of models, the methods of explanations, and outcomes of user trust. Although much of the current XAI literature has been dedicated to techniques creation, and to its critical technical functionality, there is an increasing interest in specifically applying such techniques to human experience (modalities of trust, transparency, and accountability [12], [33]) objectives.

In order to fill this gap, this paper will introduce a trust-based XAI model that has four interrelated layers:

(i) model and domain features,

(ii) XAI methodology,

(iii) property of the explanation and

(iv) user trust developments.

This multifaceted point of view allows us to have a systematic interpretation of the impact of technical design decisions on the actual adoption and confidence.

At the bottom level, the type of model and application area defines the set of XAI methods that can be feasible. As an example, tabular data are well represented using interpretable models (decision trees and generalized additive model) and deep neural networks are part of post-hoc methods of image and text processing. Essential features of the domain, like the requirement of clinical interpretability in healthcare or regulatory compliance in finance restrict even more the set of techniques used [13], [21].

The second layer is where the XAI techniques are selected which can either be intrinsic, model specific or model-agnostic. The results of each method yield such different explanations with different characteristics and trade-offs. As an example, SHAP is able to give consistent feature attribution whereas saliency maps give precedence to spatial areas in images. The decision to employ the method has to take into account the model at the level of the underlying methodology and the target of the designed methodology [39].

The third layer deals with properties of explanation namely, fidelity (faithfulness to the model), complexity (simplicity and interpretability), stability (consistency across inputs) and relevance (alignment with user context). Quality and reliability of explanations are dependent on these properties. High-fidelity explanations make sure that the users are not fooled and low complexity makes it more usable and understandable [38].

Lastly, user trust outcomes are captured in the fourth layer that comprises of perceived transparency, perceived fairness, understanding and willing to rely on the system. Trust turns out to be a task of technical and human oriented results, which makes it very important to provide the explanation that is accurate, meaningful and actionable [10], [12].

This framework underlines that explainability alone does not result in trust but rather when combined with the quality of explanations, and perception by the users. At that, to be effective, the XAI design must be done in a holistic manner incorporating both technical and human oriented aspects.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

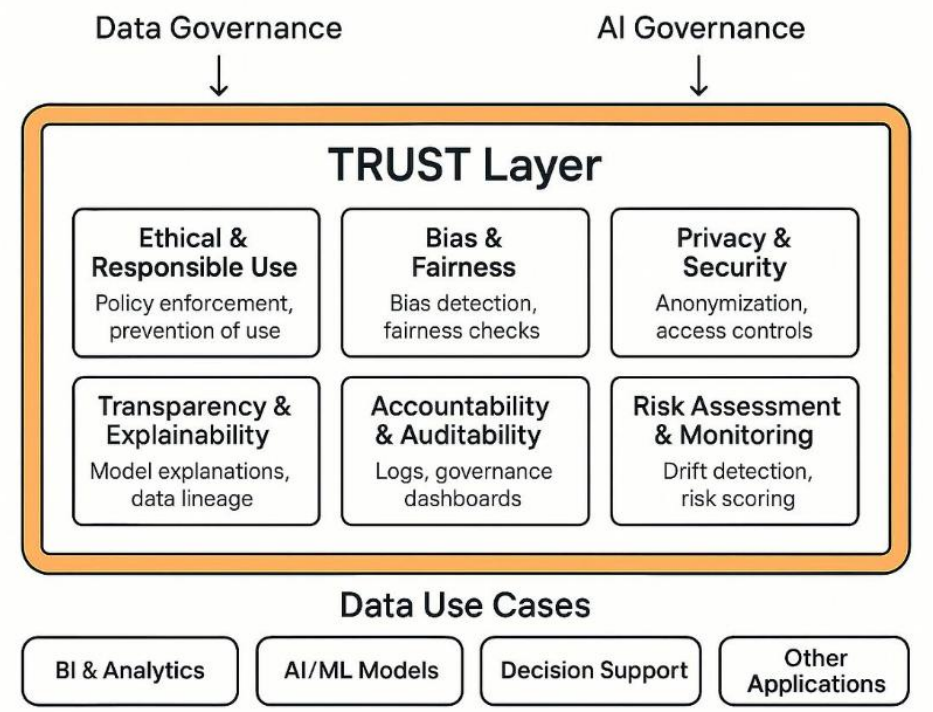

Fig. 6. Trust-centric XAI framework. An abstracted conceptual map between (i) model and data properties, (ii) choice of XAI approach, (iii) properties of an explanation e.g. fidelity and clarity, and (iv) user response e.g. trust, accountability, adoption [12], [33].

Model/Data	Domain	XAI Options	Evaluation	Benefits
Tabular	Healthcare	GAMs, SHAP	Fairness	Transparency
Image	Medical	Grad-CAM	Robustness	Visual trust
Text	Legal	Attention	Consistency	Accountability
Time-series	Finance	Attribution	Stability	Reliability
Black-box	General	LIME	Trade-off	Flexibility

TABLE V. Recommendations on the choice of XAI methods and the choice of methods depending on the type of model and the area of application and considerations of evaluation and potential benefits of trust in XAI [13], [39].

6.2 The mapping of the explanations to the needs of the stakeholders

One such component of trust-centric XAI is that explanations should be made in accordance with the requirements of various parties. As AI systems are applied by a variety of communities, merely one type of explanation might not suffice. Rather, explanations should be in terms of individual goals, expertise and expectation of each of the stakeholder groups.

- **Developers/ Data Scientists:**

The main purpose of explanations to developers is debugging, checking the models and optimization of performances. Such technical descriptions include: feature importance, gradient based methods and inside model diagnostics are key in determining when there are errors, biases or when the model is robust. In high fidelity and completeness are important since the explanations wrong can result in wrong model adjustments [13], [33].

- **Domain Experts:**

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

AI systems assist domain experts to make decisions, like clinicians, financial analysts or engineers. They need context based, trustworthy and domain knowledge related explanations. Specifically, clinicians and financial analysts need to be provided with explanations, which emphasize medically relevant characteristics and the risk factors and compliance measures, respectively. Technical productions and an example-based approach are especially helpful in the process of filling the divide between technical productions and domain knowledge [21], [40].

- **End-Users:**

The end-users who are usually the patients, customers and citizens tend to lack technical knowledge and are generally in need of straight forward, clear and practical explanations. These descriptions must also be clear reasons behind the choices; and give confidence regarding the justice and dependability of the system. In this case, counterfactual explanations and visual summaries prove to be effective as they provide information in an easy-to-understand and user friendly way [12], [33].

- **Regulators and Policymakers:**

Regulators need to have a rationale to allow an auditability assessment, transparency, and fairness. They are interested in making sure that AI systems adhere to the laws and out of legality and ethics, especially in discrimination and bias. System behavior at a policy level needs to be evaluated using global explanations, models of rules and measures of fairness. Also, the explanations should be recorded and reproduces that can undergo auditing [27], [38].

This is a stakeholder-based view that brings out the fact that trust is defined as contextual. What may work well as an explanation to a developer might not be appropriate to an end-user and the other way round. XAI systems should thus be able to provide multi-level explanations which will appeal to various audiences concurrently.

6.3 Best Practices in the selection of XAI Methods

Considering the framework suggested, it is possible to develop several top-level rules which help practitioners to choose right XAI techniques.

- **To match the method of XAI with the complexity of the data type and model used:**

Intrinsically interpretable models should be preferred (e.g., generalized additive models or sparse linear models) when possible to use tabular data due to its high stakes in such applications. In cases of the need to have black-box models, globally and locally explainable methods like SHAP should be, reliably adopted. Monotonicity and domain consistency can be used to ensure more trust can be achieved [13], [39].

- **Alcohol and Analytics:**

Fuse with other explanation styles: No one method is adequate in complicated fields like in medical imaging. Individual saliency maps can be used in conjunction with: concept-based explanations or prototype methods to give both low and high level information. Moreover, the explanations provided should be confirmed by the domain experts, to make them relevant, and accurate [20], [22].

- **Focus more on the quality, rather than the quantity of explanations:**

Giving excessive explanations on certain aspects may overpower users and diminish its usefulness. Rather, emphasize providing brief, pertinent and quality explanations that get directly to the point of task that the user is performing. This is especially needed by end-users and decision-makers that need elements of actionable insights as opposed to information that is more technical in nature [12], [33].

- **Incorporate technical assessment together with the people assessment:**

XAI methods need to be assessed both based on technical measures (e.g., fidelity, stability) and based on human measure (e.g., trust, usability). This makes sure that the explanations are not accurate, but meaningful and practical. The user studies have to be applied in high stakes situations to prove the effectiveness of explanations [38], [39].

- **Be fair, accountable and/or compliant with regulations:**

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Explanations in the regulated areas should aid in the evaluation of fairness and auditing. These involve detection of biases, recording of decisions made and adherence of the law. Easy to understand and replicate explanations are important in creating confidence with the stakeholders as well as regulators [27], [40].

- **Customize explanations to the needs of the stakeholders:**

Multi-level explanations to the various audiences should be provided by XAI systems. In the case of an example of a healthcare system, it could give detailed feature attributions to the clinicians, summaries that are simpler to the patients, as well as global reports to the regulators. Such flexibility boosts usefulness and reliability.

7 ZIPPERS, TROUBLES AND PROSPECTS

Although there have been notable improvements in the Explainable Artificial Intelligence (XAI), there are still a number of challenges which impede its reach to a broader audience and performance in application in the real world. The technical, human, and regulatory fields of concern are fluid, therefore, the need to conduct research across disciplines in order to create the genuinely trusted AI systems becomes evident.

The trade-off between usability and faithfulness has to be considered as one of the most basic issues of XAI. High fidelity explanations are a useful way to reflect internal behaviour of complex models, and are frequently very hard to interpret by people, as they are complex. On the other hand simplified explanations are simpler to grasp, but are not necessarily an accurate reflection of the logic modeled by them, and risk leading users astray [39]. Such tension brings about doubts on over-simplification whereby explanations can give intuitive but not accurate explanations that can result in misplaced trust. Striking a balance between interpretability and accuracy is an unsolved research question especially in high stakes areas where both are of paramount importance [38].

Lacking standardization with regard to XAI evaluation is another significant challenge. Even though a number of technical and human-centric measures of a quality of explanation have been put forward, no single framework of how to judge the quality of explanation exists. Various investigations would use a variety of evaluation protocols, data sets, and measures so that it would be hard to compare them and come up with benchmarks [13], [33]. This non standardness restricts the reproducibility and generalizability of XAI study. Creating a set of benchmarks and evaluation procedures as well as reports guidelines is paramount to the progress of the field and a systemic evaluation of XAI techniques.

The research/deployment gap is also also a major challenge. Although most of XAI methods have shown potential success in controlled experimental conditions, they are yet to be applied in industries. This can be explained by the fact that XAI is complicated to implement into the current systems, and there is no empirical evidence which could prove how they influence user trust and decision-making in the context of real worlds [12], [21]. To close this gap, more practical research, domain-specific validation, and larger-scale user studies to assess the efficiency of XAI techniques in operational settings are required.

More ethical and legal aspects also make the XAI systems development and deployment more complicated. XAI is commonly considered to be used in resolving such problems as bias detection, fairness, as well as accountability. Another possibility is, however, that there is superficiality in ethical washing as shallow explanations are described to present the appearance of a transparent environment without considering the biases or unethical practices that are present [27], [38]. It is a major concern to make sure that explanations are real to model behavior and can be used in promoting fairness. Moreover, such regulatory frameworks as GDPR, put an accents on the need to have explainability, although their enforcement in the reality is unclear, in particular, what qualifies as the sufficient explanation.

In the future, there are a number of exciting research directions that could be utilized to solve these problems and develop the area of XAI.

To begin with, there is an increasing trend towards causal explanations, which attempt to ground their interpretations beyond correlation-based approaches to causation and rather give insights into causation-and-effect relationships. XAI techniques may be able to provide more useful and authoritative explanations, especially in the social science sector of healthcare and policy-making by including causal reasoning [5].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Second, a direction is counterfactual explanations based on domain knowledge. Although the currently used counterfactual approaches are used to create hypothetical cases, incorporation of domain constraints can enhance the degree of realism and relevance. As an illustration, counterfactuals ought to be biologically plausible in the context of healthcare, whereas in f inance, counterfactuals ought to comply with regulatory restrictions [33].

Third, the interactive and explorational explanation systems are becoming focused upon. These systems have not offered canned explanations but instead they enable a user to query models, simulate alternate situations and dynamically hone their knowledge. These methods will be able to increase customer interaction and aid in improved decision making through live adjustment of explanations to needs of users [12]. Multi-modal XAI is another recent direction that seeks to explain a model that functions on multi-data, like text, images, and structured data. With an increasingly sophisticated AI system, the fact that it can produce consistent explanations across modalities is becoming more crucial. This applies especially to the use of autonomous systems and healthcare where the decision is made depending on a variety of data sources [11].

Additionally, with the development of foundation models and agentic AI systems, there are new challenges to explainability. Such models are very complicated and tend to work within a variety of tasks and areas thus conventional XAI is inadequate. Another key future research area will entail trying to create scalable and generalizable methods of explaining such systems [4].

Lastly, human-centered and interdisciplinary approach to XAI has to be considered. This will involve incorporating the findings of psychology, human-computer interaction and social sciences to understand better the perception and interaction of users with explanations. These techniques can assist in the development of explanations that are both technically correct and have real world significance and credibility [12], [33].

To conclude, although XAI has achieved much, the solution to these problems lies within a comprehensive strategy, which entails technical innovations, human-centered design, and regulatory conformity. Further developing XAI in these aspects will play a crucial role in creating transparent, accountable and widely trusted AI systems.

Challenge	Impact	Evidence	Future Work
Faithfulness vs Usability	Misleading	[39]	Balance
Lack of Standardization	Hard to compare	[13]	Benchmarks
Research Gap	Low adoption	[12]	Applied research
Ethical Issues	Bias risk	[27]	Fair XAI
Causal Limits	Low actionability	[5]	Causal XAI
Static Explanations	Low engagement	[12]	Adaptive XAI
Multi-modal	Complexity	[11]	Multi-modal XAI
Foundation Models	Scalability	[4]	Scalable XAI

TABLE VI. Overview of key issues impacting transparency and trust in XAI, such as faithfulness, standardization, and ethical issues, and possible research avenues [38], [39].

8 CONCLUSION

Explainable Artificial Intelligence (XAI) has risen to the stage of a critical paradigm to tackle the rising apprehensions of transparency, trust and accountability in modern machine learning (ML) systems. The requirement to have interpretable and reliable decision-making mechanisms was inevitable as AI models continue to get more and more complex and as they are applied across more high-stakes environments, including healthcare, finance, and public

policy. Black-box models can be very accurate, but fail to give insight into their internal reasoning, which restricts the adoption of these models and provokes ethical, legal, and societal worries [10], [13].

The current paper has given an analytical representation of an in-depth and organized discussion about XAI and how it can build transparency and trust. To begin with, it clarified the basic concepts, such as interpretability, explainability, transparency, and trust, and showed their connections with each other and the role they play in human-centered AI systems. Second, it has given a comprehensive taxonomy of XAI, subdividing it into intrinsically explainable models, post-hoc model specific techniques, and model-agnostic methods, thus, to provide a systematic insight into their respective applicability and trade-offs [13], [39].

Moreover, the paper synthesized both empirical and domain-specific data on the effects of explanations on trust, in various application domains. It showed that useful explanations should be contextual, brief, and be consistent with the needs of the users instead of being technical in nature. Based on these insights, a new trust-based XAI framework was introduced, which connects the properties of the models, their methods of explanation, their properties of explanation, and their outcomes in terms of trust in the user. This framework will offer a feasible guide to the design, evaluation of technical and human-centered XAI systems.

Besides, the paper outlined the main challenges and future research directions, such as the trade-off between faithfulness and usability, absence of standardized evaluation protocols, the disconnect between research and real-world deployment, and the ethical and regulatory implications of XAI [38], [39]. These issues must be tackled in order to further develop this field and provide the responsible usage of AI technologies.

To sum up, XAI is a key to making AI trustworthy as it allows connecting complex algorithms and human insight. To reach this objective, it is necessary not only to develop the correct and faithful explanation techniques, but to pay attention to the role of human factors, demands of the domain, and constraints of the regulations. A strong technical base and strict human centered design and assessment should primarily constitute effective XAI, only in such a way; it may really foster transparency, accountability, and trust in AI systems.

REFERENCES

[1] *A. Aravindkumar et al., "Explainable AI in healthcare: a systematic review of XAI applications across imaging, diagnosis, and rehabilitation," Frontiers in Artificial Intelligence, 2026, Art. no. 1749527, doi: 10.3389/frai.2026.1749527.*

[2] *V. R. Srinivas and R. Parvathi, "Explainable AI-driven MRI-based brain tumor classification: a novel deep learning approach," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1700214, 2026, doi: 10.3389/frai.2025.1700214.*

[3] *S. Dash et al., "Explainable multi-modal deep learning for transparent clinical AI," Frontiers in Artificial Intelligence, 2026, Art. no. 1767612, doi: 10.3389/frai.2026.1767612.*

[4] *S. Afroogh et al., "Beyond Explainable AI (XAI): An Overdue Paradigm Shift and Post-XAI Research Directions," arXiv preprint arXiv:2602.24176, 2026.*

[5] *A. H. Karimi et al., "Position: Explainable AI is Causality in Disguise," arXiv preprint arXiv:2603.28597, 2026.*

[6] *K. Siwek et al., "Trust Oriented Explainable AI for Fake News Detection," arXiv preprint arXiv:2603.11778, 2026.*

[7] *Y. Hahn et al., "EXplainable Classification Of DiscretE time series," arXiv preprint arXiv:2602.13087, 2026.*

[8] *A. N. M. Sakib et al., "Explainable AI for Blind and Low-Vision Users," arXiv preprint arXiv:2604.00187, 2026.*

[9] *M. Nowak et al., "Integrating Explainable AI (XAI) and NCA-Validated Interpretable Recruitment Modeling," AI, vol. 7, no. 2, Art. no. 53, 2026.*

[10] *A. Mohamed, K. Abdelqader, and K. Shaalan, "Explainable Artificial Intelligence: A systematic review of progress and challenges," Intelligent Systems with Applications, vol. 28, Art. no. 200595, 2025, doi:*

10.1016/j.iswa.2025.200595.

[11] *Z. Cheng et al., "A Comprehensive Review of Explainable Artificial Intelligence in Computer Vision," Sensors, vol. 25, no. 13, Art. no. 4166, 2025.*

[12] *H. I. Aysel et al., "Explainable Artificial Intelligence: Advancements and Future Directions," Applied Sciences, vol. 15, no. 13, Art. no. 7261, 2025.*

[13] *S. Kabir et al., "A Review of Explainable Artificial Intelligence from the Perspective of Methods, Applications, and Evaluation," Algorithms, vol. 18, no. 9, Art. no. 556, 2025.*

[14] *N. Hettikankanamage et al., "eXplainable Artificial Intelligence (XAI): A Systematic Cross Domain Review of Quantitative Prediction Tasks," Sensors, vol. 25, no. 21, Art. no. 6649, 2025.*

[15] *V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1526221, 2025, doi: 10.3389/frai.2025.1526221.*

[16] *N. Bussmann et al., "Explainable machine learning to predict the cost of capital," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1578190, 2025, doi: 10.3389/frai.2025.1578190.*

[17] *S. Hameed et al., "Explainable AI-driven depression detection from social media text," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1627078, 2025, doi: 10.3389/frai.2025.1627078.*

[18] *X. Chen et al., "Explainable machine learning to predict postoperative ileus after radical cystectomy: an 11-year real-world cohort," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1678292, 2025, doi: 10.3389/frai.2025.1678292.*

[19] *M. J. Naeen et al., "Explainable detection: a transformer-based language modeling framework for low-resource text classification," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1537432, 2025, doi: 10.3389/frai.2025.1537432.*

[20] *D. Rastogi et al., "XAI-BT-EdgeNet: explainable edge-aware deep learning for brain tumor analysis," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1676524, 2025, doi: 10.3389/frai.2025.1676524.*

[21] *R. Agrawal et al., "Integrating explainable AI (XAI) with machine learning for trustworthy healthcare diagnostics," BMC Medical Informatics and Decision Making, vol. 25, 2025, doi: 10.1186/s13000-025-01686-3.*

[22] *S. F. Nimmy, O. K. Hussain, R. K. Chakrabortty, and S. Saha, "Explainable Artificial Intelligence (XAI) in glaucoma assessment: Advancing the frontiers of machine learning algorithms," Knowledge-Based Systems, vol. 316, Art. no. 113333, 2025, doi: 10.1016/j.knosys.2025.113333.*

[23] *M. R. Shadi, H. Mirshekali, and H. R. Shaker, "Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects," Renewable and Sustainable Energy Reviews, vol. 216, Art. no. 115668, 2025, doi: 10.1016/j.rser.2025.115668.*

[24] *M. T. Ahmed, M. W. Ahmed, and M. Kamruzzaman, "A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment," Computers and Electronics in Agriculture, vol. 235, Art. no. 110354, 2025, doi: 10.1016/j.compag.2025.110354.*

[25] *I. Gómez-Talal, M. Azizsoltani, L. Bote-Curiel, J. L. Rojo-Álvarez, and A. Singh, "Towards Explainable Artificial Intelligence in Machine Learning: A study on efficient Perturbation-Based Explanations," Engineering Applications of Artificial Intelligence, vol. 155, Art. no. 110664, 2025, doi: 10.1016/j.engappai.2025.110664.*

[26] *A. Budhkar, Q. Song, J. Su, and X. Zhang, "Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics," Computational and Structural Biotechnology Journal, vol. 27, pp. 346–359, 2025, doi: 10.1016/j.csbj.2024.12.027.*

[27] *A. Emrouznejad and S. Chowdhury, "Artificial intelligence transparency and interpretability: implications*

for operations research," Annals of Operations Research, vol. 354, pp. 1–4, 2025, doi: 10.1007/s10479-025-06873-5.

[28] *G. Tzionis et al., "A review of explainable AI methods and their application in manufacturing," SN Applied Sciences, 2025, doi: 10.1007/s42452-025-07908-z.*

[29] *T. Mokheleli et al., "Explainable Artificial Intelligence for Workplace Mental Health Prediction," Information, vol. 12, no. 4, Art. no. 130, 2025.*

[30] *Y. A. Qadri et al., "Explainable Artificial Intelligence: A Perspective on Drug Discovery," Pharmaceutics, vol. 17, no. 9, Art. no. 1119, 2025.*

[31] *I. E. Agbehadji and I. C. Obagbuwa, "Explainable Artificial Intelligence and Machine Learning for Atmospheric and Environmental Prediction: A Systematic Review," Atmosphere, vol. 16, no. 10, Art. no. 1154, 2025.*

[32] *J. Moss et al., "Explainable AI in IoT: A Survey of Challenges, Techniques, and Applications," Electronics, vol. 14, no. 23, Art. no. 4622, 2025.*

[33] *X. Liu et al., "A Practical Review of Explainable Artificial Intelligence (XAI)," AI, vol. 6, no. 11, Art. no. 285, 2025.*

[34] *Z. M. Altukhi et al., "Explainable AI Definitions and Challenges in Education," arXiv preprint arXiv:2504.02910, 2025.*

[35] *L. Arora et al., "Explainable Artificial Intelligence Techniques for Software Engineering: A Survey," arXiv preprint arXiv:2505.07058, 2025.*

[36] *A. Jain et al., "Explainable AI in Big Data Fraud Detection," arXiv preprint arXiv:2512.16037, 2025.*

[37] *A. G. P. Cetina et al., "Counterfactual Explainable AI (XAI) Method for Deep Multivariate Time Series," arXiv preprint arXiv:2511.13237, 2025.*

[38] *L. Longo et al., "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," Information Fusion, vol. 106, Art. no. 102301, 2024.*

[39] *M. Mersha, K. Lam, J. Wood, A. AlShami, and J. Kalita, "Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction," arXiv preprint arXiv:2409.00265, 2024.*

[40] *K. Sadeghi R., D. Ojha, P. Kaur, R. V. Mahto, and A. Dhir, "Explainable artificial intelligence and agile decision-making in supply chain cyber resilience," Decision Support Systems, vol. 180, Art. no. 114194, 2024, doi: 10.1016/j.dss.2024.114194.*