Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Adeoye et al. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

Explainable Artificial Intelligence (XAI): A Comprehensive Review of Methods, Applications, and Open Issues

Research Article https://stem.techspherejournal.com

Article Info
Revised Date: 23rd September, 2025
Accepted Date: 25th September, 2025
Published Date: 30th September, 2025

Author Details
Adeoye Abosede Esther[1]*, Obaze Caleb Akachukwu[2]
1, 2 Department of Computer Science, Dennis Osadebay University, Asaba, Delta State, Nigeria.

****Corresponding author's email***: adeoye.esther@dou.edu.ng
DOI: https://doi.org/10.5281/zenodo.17237252

Keywords
Explainable Artificial Intelligence (XAI)
Interpretability
Trustworthy AI
Ethical AI
Human-centered AI

ABSTRACT

Artificial Intelligence (AI) has achieved remarkable breakthroughs across multiple domains, yet the increasing reliance on complex black-box models has raised concerns about trust, transparency, and accountability. Explainable Artificial Intelligence (XAI) has emerged as a critical paradigm aimed at making AI models more interpretable and understandable without compromising performance. This paper presents a comprehensive review of XAI, beginning with its foundations, historical evolution, and core principles such as interpretability, transparency, fairness, causality, and usability. It examines major methodological approaches, including model-specific versus model-agnostic techniques, intrinsic versus post-hoc explanations, and local versus global perspectives, while analyzing widely used methods such as SHAP, LIME, surrogate models, visualization tools, counterfactuals, and example-based explanations. The paper further highlights applications of XAI in healthcare, finance, autonomous systems, cybersecurity, governance, education, and recommender systems, demonstrating its relevance in real-world decision-making. Evaluation metrics including fidelity, human-centered usability, robustness, and trade-offs between explainability and performance, are discussed to frame the challenges of measuring explanation quality. Despite advancements, open issues such as lack of standardization, scalability, ethical and legal implications, and adoption barriers persist. Future directions emphasize human-centered and interactive explanations, hybrid symbolic-statistical models, standardized evaluation frameworks, applications in emerging fields, and stronger policy integration. Overall, XAI is positioned as a cornerstone for building trustworthy, sustainable, and ethical AI systems.

1 Introduction

1.1 Background of Artificial Intelligence (AI) and the Rise of Black-box Models

Artificial Intelligence (AI) has become one of the most transformative technologies of the 21st century, revolutionizing domains such as healthcare, finance, transportation, cybersecurity, and education [1]. The growing adoption of machine learning (ML) and deep learning (DL) algorithms has significantly enhanced predictive accuracy and problem-solving

capabilities [2]. However, the increasing reliance on highly complex models such as deep neural networks, ensemble methods, and large language models has introduced a critical limitation: their lack of interpretability. These models, often referred to as "black boxes," generate results without providing human-understandable reasoning or decision pathways [3]. While they excel in performance, their opacity raises concerns regarding trust, fairness, accountability, and compliance with ethical and legal standards.

1.2 Importance of Explainable AI (XAI) in Building Trust, Transparency, and Accountability

The emergence of Explainable Artificial Intelligence (XAI) addresses these limitations by focusing on methods and frameworks that make AI decisions more transparent, interpretable, and trustworthy [4]. Explainability is not merely a technical enhancement; it is a necessity for ensuring the adoption of human-centred AI. In sensitive domains such as healthcare, XAI enables physicians to understand why a diagnostic model suggests a particular treatment, thereby improving clinical confidence and patient safety [5]. In finance, it ensures that credit scoring and fraud detection models comply with regulatory requirements and ethical considerations. Furthermore, XAI plays a crucial role in enhancing user trust, enabling meaningful human-AI collaboration, mitigating bias, and ensuring accountability in decision-making processes [6]. Without explainability, AI systems risk rejection, misuse, or unintended harmful consequences.

1.3 Objectives of the Review Paper

This review paper aims to provide a comprehensive overview of the current state of Explainable AI by:

a. Presenting the conceptual foundations and principles of XAI.
b. Reviewing major methods and approaches used to enhance interpretability, including model-specific and model-agnostic techniques.
c. Highlighting practical applications of XAI across diverse domains such as healthcare, finance, cybersecurity, autonomous systems, and governance.
d. Discussing evaluation metrics for explainability and the trade-offs between performance and interpretability.
e. Identifying open challenges, limitations, and research gaps that hinder widespread adoption of XAI.
f. Suggesting future research directions to advance trustworthy, transparent, and ethically aligned AI systems.

1.4 Structure of the Paper

The remainder of this paper is structured as follows:

a. **Section 2** introduces the foundations of XAI, including its definitions, historical development, and key principles.
b. **Section 3** reviews prominent methods and approaches for explainability, categorizing them into intrinsic and post-hoc, as well as local and global techniques.
c. **Section 4** explores real-world applications of XAI across multiple industries.
d. **Section 5** presents evaluation metrics for assessing the quality and effectiveness of explanations.
e. **Section 6** discusses open challenges, ethical concerns, and unresolved issues in the adoption of XAI.
f. **Section 7** highlights future directions and emerging trends in XAI research.
g. **Section 8** concludes the paper with key insights and closing remarks.

2 Foundations of Explainable Artificial Intelligence

2.1 Definition and Scope of XAI

Explainable Artificial Intelligence (XAI) refers to a set of methods, techniques, and frameworks designed to make the decision-making processes of AI systems more transparent, understandable, and interpretable to human users [7]. Unlike traditional "black-box" AI models, where outputs are generated without insight into the reasoning process, XAI aims

to provide human-interpretable explanations that justify model predictions or decisions [8]. The scope of XAI extends beyond technical transparency; it also encompasses building trust, ensuring fairness, facilitating accountability, and enabling meaningful human-AI interaction [9]. Depending on the application, explanations may be required for end-users (to understand system behavior), developers (to debug and optimize models), or regulators (to ensure compliance with ethical and legal standards). Thus, XAI serves as both a technical and socio-ethical framework that bridges advanced AI capabilities with responsible and trustworthy usage [10].

2.2 Evolution and Historical Context

The need for explainability in AI is not new. Early expert systems in the 1970s and 1980s, such as MYCIN in medical diagnosis, incorporated rule-based reasoning that was inherently interpretable, allowing users to trace back the logic behind decisions [11]. With the rise of machine learning in the 1990s and the dominance of deep learning in the 2010s, the complexity of models increased dramatically, leading to a shift from interpretable symbolic reasoning to opaque statistical learning [12]. This opacity raised concerns about accountability and ethical use, particularly in high-stakes domains like healthcare, finance, and criminal justice. The introduction of landmark regulations such as the European Union's General Data Protection Regulation (GDPR) in 2018, which emphasized a "right to explanation," further accelerated research into XAI [13], [14], [15]. Today, XAI is a rapidly expanding research field, blending contributions from computer science, cognitive psychology, human-computer interaction, philosophy, and law [16].

2.3 Core Principles of Explainability

Several fundamental principles guide the development and evaluation of XAI systems:

1. **Interpretability:** The degree to which a human can understand the internal mechanics of a model. For example, decision trees and linear regression models are highly interpretable compared to deep neural networks [5]
2. **Transparency:** Refers to the openness of AI models in terms of their design, data usage, and decision-making process. Transparent models allow stakeholders to trace inputs to outputs [17].
3. **Fairness:** Ensures that AI models provide equitable outcomes without reinforcing biases or discrimination. Explanations can reveal hidden biases in data and algorithms, making fairness a key principle of XAI.
4. **Causality:** Goes beyond correlation-based explanations to highlight cause-and-effect relationships. Causal reasoning enhances trust by showing how specific features or inputs directly influence outcomes [18].
5. **Usability:** Explanations should be designed in a way that is accessible and meaningful to their intended audience. A clinician, policymaker, or lay user may require different forms of explanation for the same AI decision [19].

These principles collectively ensure that AI systems are not only powerful but also human-centric and socially responsible.

2.4 Relationship between Explainability and AI Ethics

Explainability is a cornerstone of ethical and trustworthy AI. Ethical concerns in AI, such as bias, discrimination, privacy invasion, and lack of accountability, are often exacerbated by opaque decision-making [20]. XAI mitigates these issues by providing clarity into how decisions are made, thereby enabling oversight and responsibility [21]. For instance, in healthcare, explainable diagnostic systems support the ethical principle of beneficence by ensuring that treatments are evidence-based and understandable to both doctors and patients [22]. In finance, explainability supports the principle of justice, ensuring fairness in credit scoring and lending practices [23]. Moreover, explainability strengthens accountability, allowing institutions and developers to justify and defend AI decisions to regulators and society [24]. By aligning technical transparency with ethical principles, XAI fosters responsible innovation, ensuring that AI systems serve humanity in a fair, safe, and trustworthy manner [25].

Table 1 and Figure 1 present the core principles and foundations of XAI, while Table 2 presents the historical evolution.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Adeoye et al. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

Table 1: Core Principles of Explainable AI

Principle	Definition	Practical Example in XAI
Interpretability	The degree to which a human can understand the internal mechanics of an AI system or model output.	A linear regression model in healthcare predicting blood pressure from age and weight—where coefficients directly indicate the relationship between variables.
Transparency	The openness of an AI system in exposing its decision-making process and structure.	A decision tree used in finance for credit scoring, where each branching rule is visible and explainable.
Fairness	Ensuring that AI systems do not produce biased or discriminatory outcomes across different demographic groups.	An algorithm for hiring that is checked for bias by comparing outcomes across gender or ethnic categories.
Causality	Providing explanations that reveal cause-and-effect relationships rather than mere correlations.	A medical AI system that explains how smoking increases the risk of lung cancer, rather than just showing an association.
Usability	The extent to which explanations are actionable, user-friendly, and aligned with the needs of stakeholders.	An AI-based recommender system in e-commerce that explains "You were shown this product because of your previous purchases."

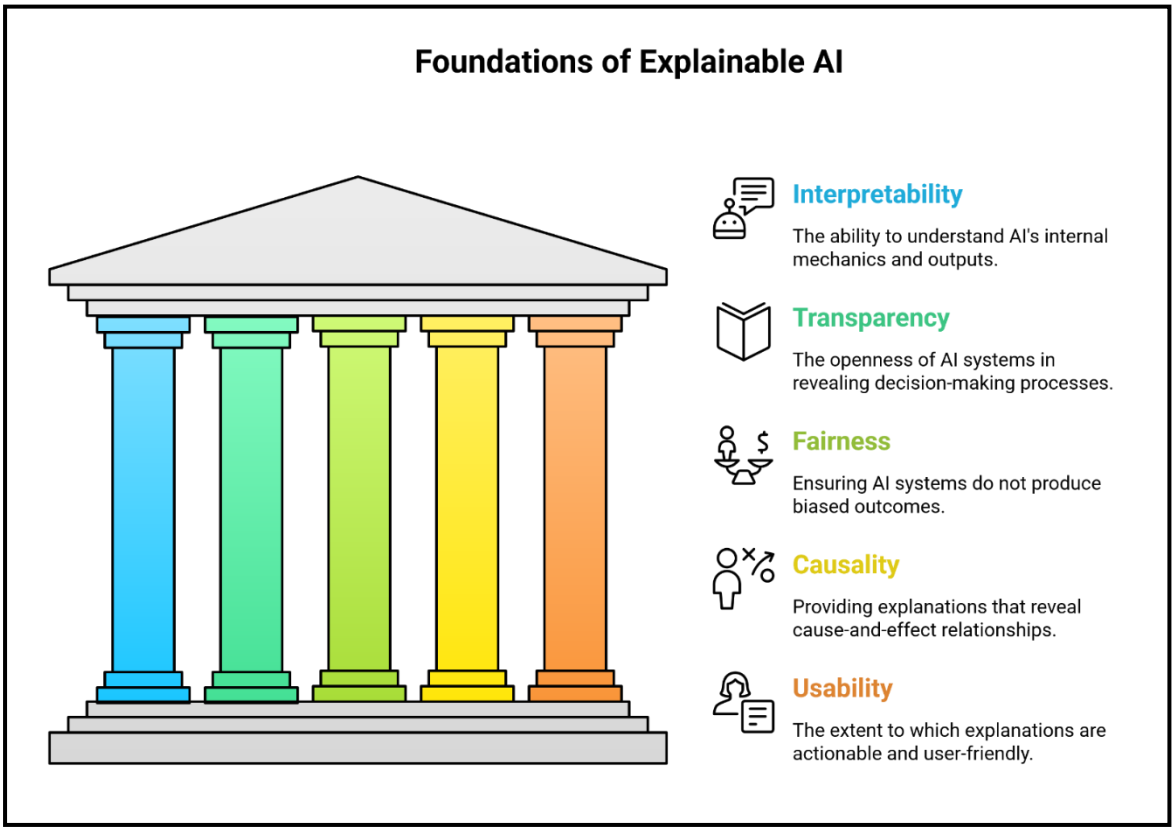

Figure 1: The Foundations of XAI

Table 2: Historical Evolution of Explainable AI (XAI)

Era / Period	Key Characteristics	Milestones / Examples	Limitations / Challenges
1970s–1980sExpert Systems Era	• Rule-based AI• Logic-driven reasoning• Human-readable explanations	MYCIN, DENDRAL, rule-based expert systems	Limited scalability, brittle rules, poor generalization
1990s Machine Learning Emergence	• Transition to statistical models• Data-driven learning	Neural networks, support vector machines (SVMs)	Reduced interpretability, rise of "black-box" models
2000s Early Interpretability Tools	• Renewed interest in interpretability• Feature selection & decision trees	CART, Random Forest feature importance, PCA	Limited to shallow models; insufficient for deep architectures
2010sDeep Learning Dominance	• Widespread adoption of deep neural networks• Opaque and complex models	CNNs, RNNs, reinforcement learning	High predictive power but "black-box" nature raised trust issues
2017+Dedicated XAI Methods	• Birth of XAI as a research field• Post-hoc model explanation	LIME, SHAP, Grad-CAM, counterfactual explanations	Focus mainly on local explanations; issues of consistency
2020sResponsible & Trustworthy AI	• Integration of ethics and fairness• Accountability & transparency frameworks	DARPA XAI program, EU AI Act, Explainable ML pipelines	Balancing performance with fairness and interpretability

3 Methods and Approaches in Explainable Artificial Intelligence (XAI)

Explainable AI has evolved into a diverse field, with multiple approaches designed to enhance transparency and interpretability across different types of models [26]. This section examines the conceptual foundations of XAI methods and provides an overview of the techniques commonly employed.

3.1 Model-specific vs. Model-agnostic Approaches

a. **Model-specific approaches** are tailored to a particular class of models and leverage their internal structure to generate explanations. For instance, decision trees naturally yield human-readable rules, while convolutional neural networks (CNNs) can be explained through saliency maps that highlight influential pixels [27]. These methods often provide more accurate explanations but are limited to the specific model type chosen.

b. **Model-agnostic approaches**, in contrast, treat the model as a "black box" and provide explanations without requiring access to its internals [28]. Techniques such as LIME and SHAP fall into this category, as they approximate the decision boundary using simpler interpretable models. Their versatility makes them broadly applicable, but the fidelity of the explanation to the original model may vary.

3.2 Post-hoc vs. Intrinsic Explainability

a. **Intrinsic explainability** refers to models that are inherently interpretable by design. Examples include linear regression, decision trees, and rule-based systems, which present their logic in a form directly understandable to humans [29]. The advantage lies in simplicity and direct interpretability, but these models may struggle to capture highly complex patterns in data.

b. **Post-hoc explainability**, on the other hand, involves applying external techniques to interpret models after training [30]. Deep neural networks, for example, are often explained post-hoc through feature attribution or visualization techniques. While this approach allows the use of high-performing complex models, the explanations are approximate and sometimes inconsistent.

3.3 Local vs. Global Explanations

a. **Local explanations** aim to clarify individual predictions by showing which features influenced a specific outcome [31]. LIME and SHAP are widely used for local interpretability, helping users understand "why this decision was made."
b. **Global explanations** focus on the overall logic and behaviour of the model across the dataset. Examples include partial dependence plots, feature importance rankings, and interpretable surrogate models that approximate the full decision boundary [32]. Global methods are essential for understanding systematic biases and fairness considerations.

3.4 Common XAI Techniques

1. Feature Attribution Methods

These techniques quantify the contribution of each input feature to a model's prediction:

a. **LIME (Local Interpretable Model-agnostic Explanations):** Creates local surrogate models around an instance to approximate its decision boundary [33].
b. **SHAP (SHapley Additive exPlanations):** Based on Shapley values from cooperative game theory, providing consistent feature importance values [34].

2. Visualization Methods

Visual representations make the inner workings of models more accessible:

a. **Saliency maps and Grad-CAM:** Highlight regions of input (e.g., image pixels) most influential in the prediction [35].
b. **Partial dependence plots (PDPs):** Show how a feature impacts predictions across its range, helping reveal non-linear relationships [36].

3. Surrogate Models

Interpretable models trained to approximate complex models:

a. Decision trees and rule-based systems can act as simplified stand-ins for black-box models.
b. Useful for global understanding but may oversimplify complex behaviors.

4. Counterfactual Explanations

These methods answer the question*: "What minimal changes in input features would flip the model's decision?"* For example, a loan rejection could be explained by showing that an increase in annual income by $5,000 would have led to approval. Counterfactuals are highly intuitive for end-users but may be computationally intensive to generate.

Example-based Explanations

Models can be explained by presenting representative or influential examples:

a. Prototype-based methods identify instances most representative of a class.
b. Criticism-based methods highlight atypical cases that deviate from the norm.

This approach aligns well with human reasoning, as people often understand concepts through examples. Table 3 presents the comparative evaluation of XAI methods and approaches.

Table 3: Comparative Evaluation of XAI Methods and Approaches

Approach / Method	Type	Strengths	Weaknesses	Examples / Techniques
Model-Specific	Intrinsic / Post-hoc	High fidelity to the chosen model; leverages internal structure	Limited to a particular algorithm family	Saliency maps for CNNs, attention visualization in Transformers
Model-Agnostic	Post-hoc	Flexible, applicable to any model; provides general interpretability	May approximate poorly; computationally expensive	LIME, SHAP
Intrinsic Explainability	Built-in	Simple, human-readable; no extra processing needed	Struggles with complex data; may sacrifice accuracy	Linear regression, Decision trees, Rule-based systems
Post-hoc Explainability	Add-on	Enables interpretability for high-performing black-box models	Explanations approximate; potential inconsistencies	Feature attribution, Visualization, Surrogate models
Local Explanations	Instance-level	Explains individual predictions; high relevance to users	May not generalize globally; possible instability	LIME, SHAP (local), Counterfactuals
Global Explanations	Model-level	Provides overall model behavior insights; helps detect bias	May miss nuances of individual predictions	Feature importance, PDPs, Surrogate trees
Feature Attribution	Post-hoc	Clear feature contribution scores; widely adopted	Sensitive to feature correlations; computational cost	SHAP, LIME, Integrated Gradients
Visualization Methods	Post-hoc	Intuitive for humans; helps in image/NLP tasks	Limited to specific data types; may oversimplify	Saliency maps, Grad-CAM, PDPs
Surrogate Models	Post-hoc (Global)	Simplifies black-box into interpretable form	Risk of oversimplification; reduced fidelity	Decision tree surrogates, Rule extraction
Counterfactual Explanations	Post-hoc (Local)	Intuitive "what-if" reasoning; actionable insights	Computationally intensive; feasibility constraints	Loan approval "what-if" scenarios
Example-based Explanations	Post-hoc (Local/Global)	Aligns with human reasoning; highlights prototypes & outliers	Depends on quality of dataset; may lack generalization	Prototype selection, Criticism detection

4 Applications of Explainable AI

The practical significance of Explainable AI (XAI) lies in its ability to make opaque AI systems more transparent, interpretable, and trustworthy across different sectors. By bridging the gap between complex models and human understanding, XAI not only enhances user confidence but also ensures ethical compliance, accountability, and effective decision-making. This section highlights the major application domains of XAI.

4.1 Healthcare and Medical Diagnostics

AI is widely used in healthcare for disease prediction, medical imaging, drug discovery, and personalized treatment recommendations [37]. However, due to the high-risk nature of medical decisions, explainability is critical. XAI helps physicians understand why an algorithm suggested a diagnosis (e.g., highlighting tumor regions in an MRI) or why a patient was classified as high-risk [38]. Techniques like saliency maps, SHAP values, and counterfactual explanations ensure that clinicians can trust the system's recommendations while retaining responsibility for final decisions.

4.2 Finance and Banking (Fraud Detection, Credit Scoring)

Financial institutions employ AI for credit risk assessment, loan approvals, fraud detection, and algorithmic trading. XAI ensures transparency in these processes, particularly in regulatory contexts such as GDPR and Basel III [39]. For instance, in credit scoring, XAI can provide actionable explanations (e.g., "loan denied due to insufficient income or high debt ratio"), which prevents discrimination and supports fairness. Similarly, explainable fraud detection systems allow auditors and analysts to trace suspicious activities instead of treating model outcomes as black boxes.

4.3 Autonomous Systems (Self-driving Cars, Robotics)

Autonomous vehicles and intelligent robots rely on deep learning models for object recognition, decision-making, and navigation [40]. However, safety-critical operations demand transparency in their reasoning. XAI allows engineers and regulators to investigate why a car decided to brake, change lanes, or misinterpret a pedestrian crossing [41]. Techniques like visual saliency, interpretable sensor fusion, and rule-based surrogates can help validate these systems and prevent catastrophic failures.

4.4 Cybersecurity (Intrusion Detection, Anomaly Detection)

AI models are increasingly applied in detecting cyber threats, phishing attacks, and anomalies in network traffic. Yet, black-box predictions can be difficult for security analysts to act upon. XAI provides interpretable alerts by highlighting which features of a data packet or user behavior led to its classification as malicious [42]. This transparency accelerates incident response, reduces false positives, and builds trust in AI-driven security platforms.

4.5 Legal, Governance, and Policy Domains

In law and governance, decision-making often involves sensitive issues of fairness, accountability, and rights. XAI is vital for ensuring that algorithmic decisions in domains such as judicial risk assessment, parole recommendations, or welfare eligibility are interpretable and non-discriminatory [43]. Policymakers can rely on XAI to audit AI systems, ensuring compliance with ethical frameworks and legal standards. Moreover, explainability helps balance automation with democratic oversight in governance.

4.6 Education, Human-Computer Interaction, and Recommender Systems

XAI is transforming education and learning platforms by making AI-driven feedback transparent to both learners and instructors. For example, intelligent tutoring systems that explain why a student received a recommendation for extra practice foster trust and personalized learning. In recommender systems (e.g., e-commerce or streaming platforms), XAI provides users with rationales for recommendations ("You were shown this movie because you liked similar titles"), thereby increasing engagement and reducing perceptions of bias [44]. Similarly, in human-computer interaction, explainable systems improve usability and foster trust between users and AI agents.
Table 4 and Figure 2 present the applications of XAI across domains.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Adeoye et al. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

Table 4: Applications of XAI across domains

Domain	Role of XAI	Key Techniques Commonly Used
Healthcare & Medical Diagnostics	Improves trust in diagnostic AI, supports clinical decisions, highlights critical medical features (e.g., tumors, biomarkers).	Saliency maps, SHAP values, counterfactual explanations, partial dependence plots.
Finance & Banking	Ensures fairness in credit scoring, enables transparent loan approvals, enhances fraud detection interpretability.	LIME, SHAP, rule-based surrogate models, feature attribution.
Autonomous Systems (Cars & Robotics)	Explains navigation and safety decisions, validates model outputs, prevents catastrophic failures.	Visual saliency maps, interpretable sensor fusion, decision trees as surrogates.
Cybersecurity (Intrusion & Anomaly Detection)	Provides interpretable alerts, reduces false positives, assists analysts in incident response.	Feature attribution, rule-based models, anomaly score explanations.
Legal, Governance & Policy	Promotes fairness and accountability, ensures compliance with ethical and legal standards, supports algorithmic auditing.	Rule-based systems, interpretable decision trees, counterfactual analysis.
Education, HCI & Recommender Systems	Enhances learner trust, explains personalized feedback, provides rationale for recommendations, improves user experience.	Example-based explanations, transparent recommendation rationales, surrogate models.

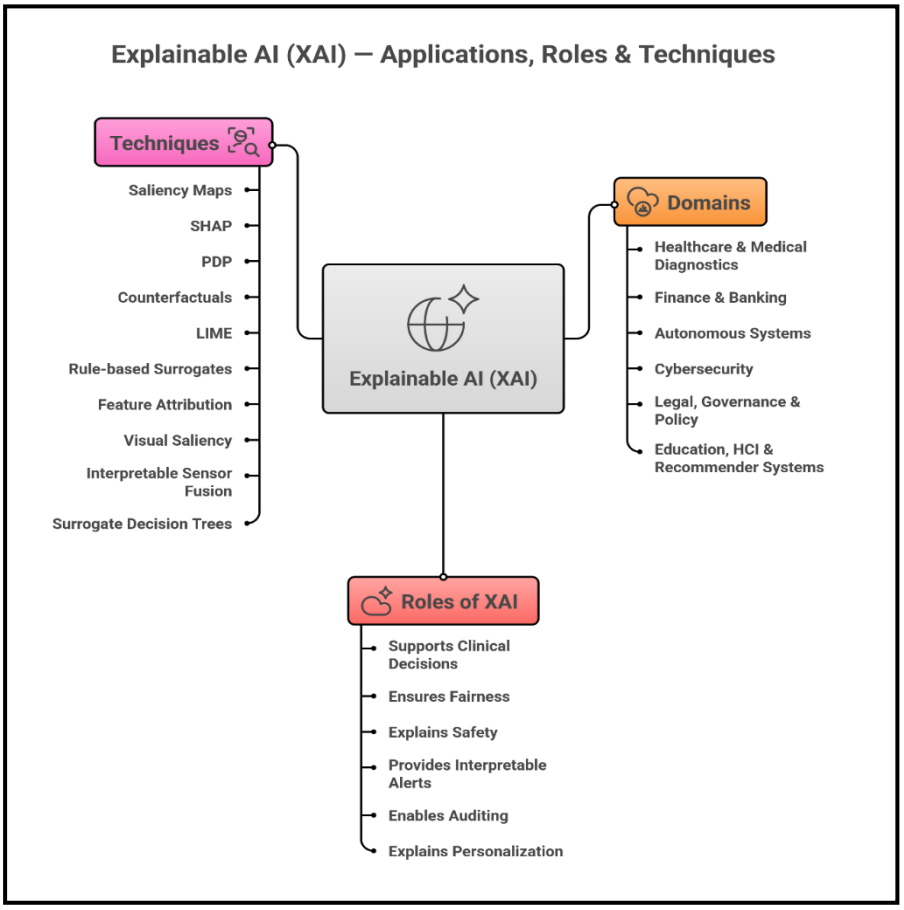

Figure 2: Applications and Roles and Techniques of XAI

5 Evaluation Metrics for Explainability

The effectiveness of XAI systems depends not only on their ability to generate explanations but also on how these explanations are evaluated. Unlike traditional machine learning models, which are typically assessed using predictive accuracy or error rates, XAI requires a broader and multidimensional evaluation framework. Such metrics capture both technical validity and human-centered impact, ensuring that explanations are not only faithful to the underlying model but also understandable and actionable for end users. This section explores key evaluation dimensions.

5.1 Fidelity and Accuracy of Explanations

Fidelity refers to the degree to which an explanation accurately reflects the internal reasoning of the underlying model [45]. An explanation is considered high-fidelity if the simplified interpretation (e.g., surrogate model, feature attribution) closely approximates the predictions of the original model. Fidelity can be quantitatively measured by comparing the outputs of the interpretable explanation method against the outputs of the complex model using statistical metrics such **as correlation coefficients, mean squared error (MSE), or classification agreement rates**.

a. **Example:** A surrogate decision tree approximating a deep neural network should produce predictions that align with the network in most cases.
b. **Challenge:** High fidelity does not necessarily imply comprehensibility to humans, underscoring the need for balancing technical and user-centered measures.

5.2 Human-Centered Metrics (Usability, Comprehensibility, Trust)

Human-centred metrics assess how effectively explanations serve their ultimate audience: people [46]. Explanations must be comprehensible, usable, and trustworthy for stakeholders, including domain experts, regulators, and end users.

a. **Usability** can be measured through task performance, e.g., whether users can make faster or more accurate decisions with the aid of explanations.
b. **Comprehensibility** can be evaluated through user studies, measuring subjective ratings of clarity, simplicity, or ease of understanding.
c. **Trust** reflects the confidence users place in an AI system after receiving explanations, often measured via surveys, Likert-scale questionnaires, or behavioral experiments.
d. These metrics highlight the socio-technical dimension of XAI, ensuring that explanations are not only algorithmically valid but also contextually meaningful.

5.3 Robustness and Consistency

One of the most critical evaluation considerations is the trade-off between model performance and interpretability. Highly accurate models such as deep neural networks often operate as "black boxes," while simpler models (e.g., linear regression, decision trees) offer greater transparency at the cost of reduced predictive accuracy [47].

a. In some domains (e.g., healthcare, finance, law), interpretability may be prioritized over raw accuracy due to regulatory or ethical concerns.
b. Conversely, in high-performance domains (e.g., computer vision), accuracy may be critical, and post-hoc explanation techniques are used to strike a balance.
c. Evaluating this trade-off involves multi-objective metrics, balancing predictive performance (e.g., accuracy, precision, recall) with interpretability scores (e.g., simplicity of rules, explanation length, user ratings). This requires careful consideration of context-specific priorities and stakeholder needs.

Table 5 and Figure 3 present the comparative summary of evaluation metrics for XAI

Table 5: Comparative Summary of Evaluation Metrics for Explainability

Metric	Definition	Examples of Measurement	Key Advantages	Challenges/Limitations
Fidelity & Accuracy	The degree to which explanations reflect the true reasoning of the underlying model.	Correlation coefficients, Mean Squared Error (MSE) between surrogate and original model, Classification agreement rates.	Ensures explanations are technically valid; provides quantitative benchmarks.	High fidelity does not guarantee human comprehensibility; may still be opaque to non-experts.
Human-Centered Metrics (Usability, Comprehensibility, Trust)	Extent to which explanations are useful, understandable, and inspire confidence among end users.	User studies, Task performance improvement, Likert-scale surveys, Comprehension tests.	Captures real-world effectiveness and user satisfaction; ensures socio-technical alignment.	Subjective and context-dependent; difficult to standardize across domains.
Robustness & Consistency	Stability and reliability of explanations under small perturbations or across methods.	Perturbation tests, Stability indices, Cross-method agreement (e.g., SHAP vs. LIME results).	Ensures explanations are reliable and reproducible; improves trustworthiness.	Computationally expensive to evaluate; conflicting outputs may confuse users.
Trade-off between Explainability & Performance	Balancing interpretability of the model with predictive accuracy.	Comparing accuracy/precision with interpretability metrics (e.g., rule length, explanation complexity).	Provides context-specific balance between accuracy and transparency; useful for regulatory compliance.	No universal balance point; priorities differ by domain (e.g., medicine vs. image recognition).

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Adeoye et al. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

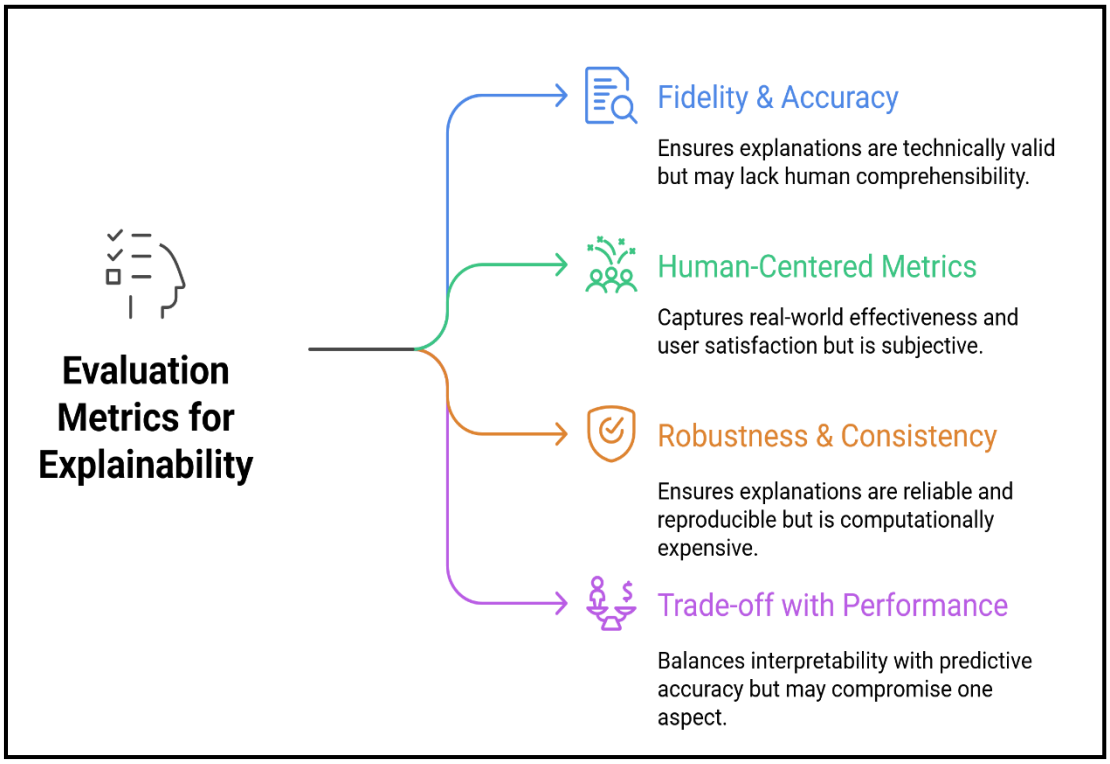

Figure 3: Evaluation Metrics for Explainable AI

6 Open Issues and Challenges in XAI

Despite significant progress in Explainable Artificial Intelligence (XAI), several open issues and challenges hinder its widespread adoption and effectiveness. These challenges span technical, methodological, ethical, and industrial dimensions, highlighting the need for interdisciplinary approaches to advance the field.

6.1 Lack of Standardization in Metrics and Benchmarks

Currently, there is no universally accepted framework for evaluating XAI methods [48]. Different studies rely on diverse metrics such as fidelity, comprehensibility, or user trust, making it difficult to compare models and explanations across domains. The absence of standardized benchmarks also limits reproducibility and consistency in research findings. Establishing a global set of evaluation criteria and datasets would facilitate fair comparisons, accelerate innovation, and foster transparency in XAI research.

6.2 Balancing Accuracy vs. Interpretability

A major challenge in XAI is the inherent trade-off between model complexity and interpretability [49]. Highly accurate models, such as deep neural networks, often produce opaque predictions, whereas simpler models (e.g., linear regression, decision trees) are easier to explain but may sacrifice accuracy. Designing hybrid or layered models that balance predictive performance with human interpretability remains an open research direction.

6.3 Scalability of XAI Methods

Most current XAI methods struggle to scale effectively to real-world, large-scale datasets and high-dimensional problems. For instance, generating explanations for deep learning models applied to multimodal or streaming data can

be computationally expensive. Achieving scalable, efficient, and real-time explainability remains a pressing challenge, especially in domains like finance, healthcare, and autonomous systems where decisions are time-sensitive.

6.4 Ethical, Legal, and Social Implications (ELSI)

XAI extends beyond technical performance to encompass broader societal considerations. Ethical issues include the potential misuse of explanations to manipulate users, privacy concerns when explanations reveal sensitive information, and accountability in cases of system failure [7]. From a legal standpoint, regulations such as the EU's General Data Protection Regulation (GDPR) mandate the "right to explanation," raising questions about compliance and enforceability. Addressing these implications requires collaboration between computer scientists, ethicists, policymakers, and legal experts.

6.5 Bias, Fairness, and Accountability in Explanations

Explanations themselves may introduce or perpetuate biases, especially when generated from biased data or model structures. For example, an explanation system may justify discriminatory decisions in hiring, lending, or law enforcement. Ensuring fairness, accountability, and transparency in XAI requires rigorous bias detection mechanisms, debiasing techniques, and frameworks for auditing explanations [50]. Without such safeguards, XAI could inadvertently reinforce systemic inequities.

6.6 Adoption Barriers in Industry

Despite growing academic interest, the industrial adoption of XAI remains limited. Barriers include the lack of trust in explanation fidelity, integration challenges with existing workflows, concerns over computational overhead, and insufficient regulatory clarity [51]. Additionally, many organizations struggle to balance business incentives with transparency initiatives. Bridging the gap between research prototypes and deployable solutions demands greater collaboration between academia, industry stakeholders, and regulatory bodies.
Table 5 and Figure 4 present the open issues and challenges in XAI

Table 6: Open Issues and Challenges in XAI

Challenge	Key Issue	Future Research Direction
Lack of Standardization in Metrics and Benchmarks	No unified metrics or benchmarks to evaluate XAI methods; inconsistency across studies	Develop standardized evaluation frameworks, shared datasets, and reproducible benchmarks for fair comparison
Balancing Accuracy vs. Interpretability	Trade-off between model complexity and human comprehensibility	Explore hybrid models, modular interpretability layers, and adaptive explanation systems
Scalability of XAI Methods	High computational cost and limited scalability to large-scale or real-time applications	Design efficient, scalable algorithms for multimodal and streaming data in dynamic environments
Ethical, Legal, and Social Implications (ELSI)	Concerns over privacy, misuse of explanations, accountability, and regulatory compliance	Develop privacy-preserving explainability methods; interdisciplinary collaboration with ethicists, legal scholars, and policymakers
Bias, Fairness, and Accountability in Explanations	Risk of perpetuating or justifying discriminatory outcomes	Create auditing frameworks, bias detection and debiasing tools, and fairness-aware explanation models
Adoption Barriers in Industry	Limited trust, integration challenges, and unclear ROI hinder real-world deployment	Develop user-centered XAI tools, industry standards, and regulatory guidance to support practical adoption

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Adeoye et al. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

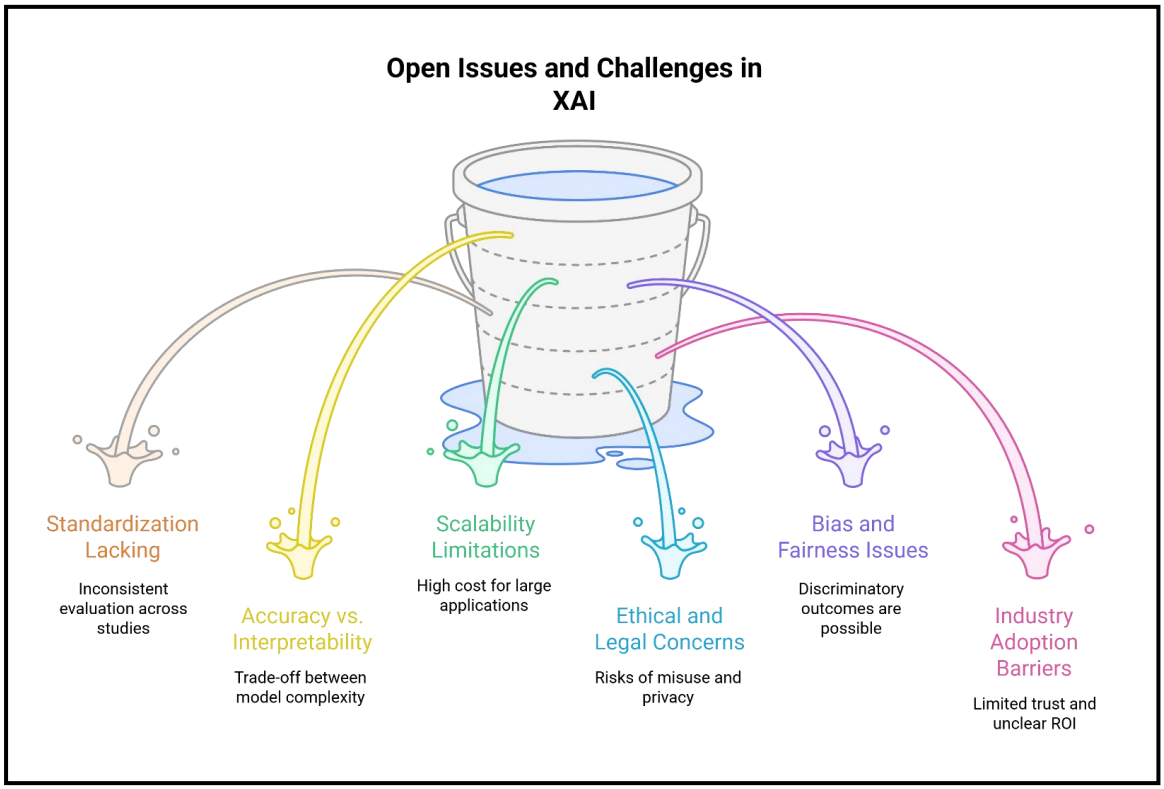

Figure 4 Open Issues and Challenges in XAI

7 Future Directions in XAI Research

7.1 Human-Centered and Interactive Explanations

The next generation of XAI must prioritize human-centered design, focusing on explanations that are not only technically accurate but also usable, contextual, and adaptive to different stakeholders (e.g., clinicians, regulators, end-users). Interactive XAI systems, where users can query, refine, or challenge explanations, will foster trust and promote continuous learning between humans and AI [52]. Such systems can personalize explanations based on user expertise and preferences, bridging the gap between transparency and comprehension.

7.2 Hybrid Approaches (Combining Symbolic and Statistical AI)

A promising research direction is the integration of symbolic reasoning (e.g., logic, rules, knowledge graphs) with statistical AI (e.g., deep learning) [53]. Hybrid models aim to achieve both high predictive accuracy and interpretability, leveraging the strengths of each paradigm. Symbolic layers can encode domain knowledge and constraints, while statistical components can capture complex patterns. This synergy supports faithful and interpretable reasoning pipelines, particularly in domains such as healthcare, law, and autonomous systems.

7.3 Standard Frameworks for Evaluation

The absence of universally accepted benchmarks remains a critical challenge. Future research must develop standardized evaluation frameworks to compare XAI methods on common datasets, tasks, and metricsv [54]. Such frameworks should account for technical fidelity, human usability, robustness, and fairness, enabling reproducibility and consistency

across studies. Establishing community-driven benchmarks will ensure progress is comparable, transparent, and scientifically rigorous.

7.4 Explainability in Emerging Fields (Federated Learning, Generative AI, Edge AI)

As AI expands into new paradigms, XAI must evolve accordingly:

a. **Federated Learning:** Explanations must respect privacy constraints while clarifying model behavior in decentralized settings.
b. **Generative AI:** Explaining generative processes (e.g., large language models, diffusion models) requires novel tools to capture both latent space reasoning and content plausibility.
c. **Edge AI:** Lightweight, real-time explanations are crucial for resource-constrained devices (e.g., wearables, IoT sensors), balancing interpretability with latency and energy efficiency.

7.5 Policy and Regulatory Implications

Growing adoption of AI in sensitive domains will necessitate policy frameworks and regulatory standards **a**round explainability. Legislations such as the EU's *AI Act* and global governance initiatives will likely mandate transparency, fairness, and accountability in AI systems [55]. Research should explore how to align XAI methods with legal requirements while maintaining usability and technical rigor. Policymakers, technologists, and ethicists must collaborate to ensure explainability serves as a safeguard against misuse and fosters **responsible AI deployment**.

Figure X presents the conceptual roadmap of XAI.

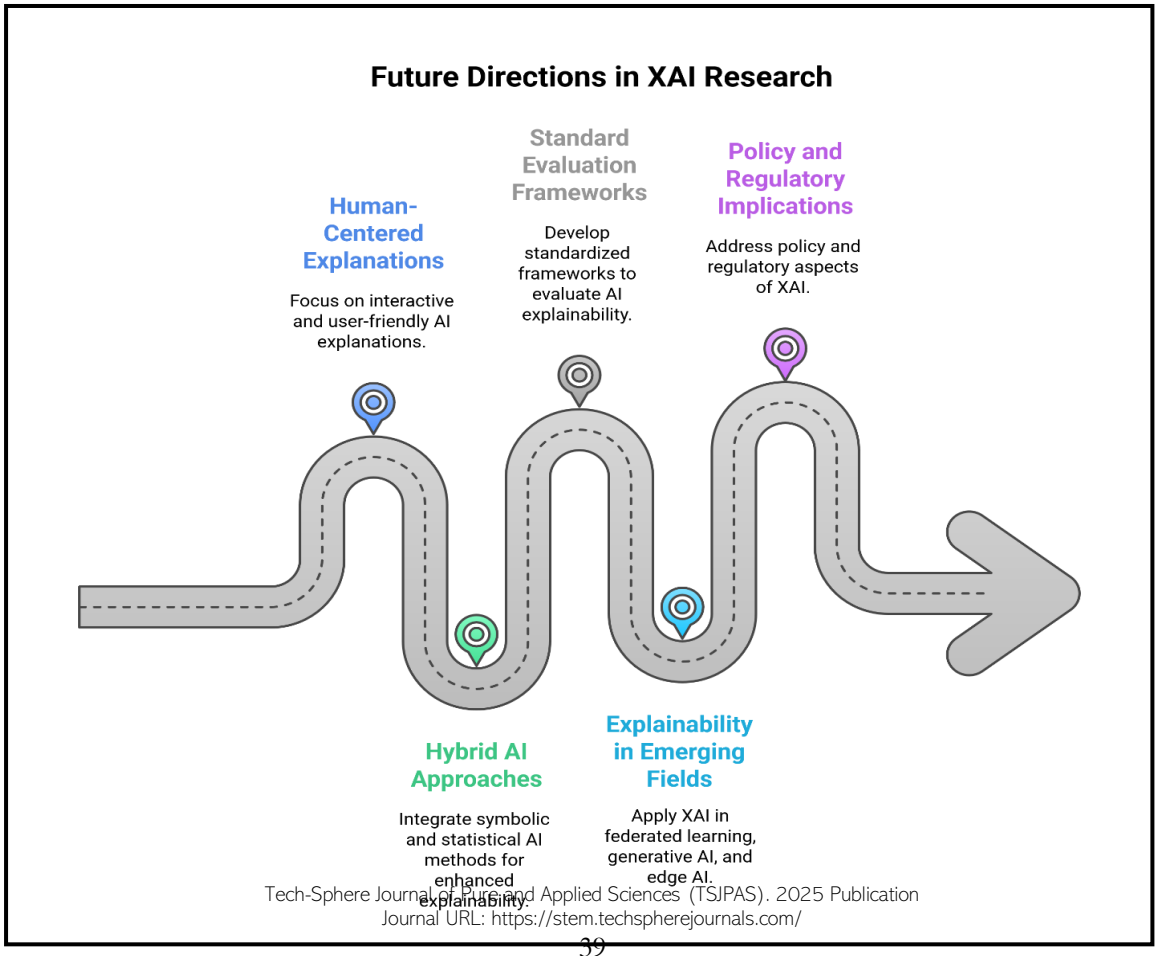

Figure 5: Future Directions of XAI Research

8 Conclusion and Future Work

Explainable Artificial Intelligence (XAI) has emerged as a critical paradigm in addressing the inherent opacity of modern AI systems, particularly those dominated by complex deep learning architectures. This review has presented a comprehensive exploration of the foundations, methods, applications, and evaluation metrics of XAI, alongside the open issues and future research directions that continue to shape the field. From a methodological standpoint, both model-specific and model-agnostic approaches, post-hoc and intrinsic techniques, and local versus global explanations provide researchers and practitioners with a diverse toolkit for addressing the explainability challenge. Techniques such as feature attribution, visualization methods, surrogate models, counterfactual reasoning, and example-based explanations demonstrate that there is no one-size-fits-all solution, but rather a spectrum of approaches suited to different contexts. The review also highlighted how XAI is increasingly integrated into critical applications such as healthcare, finance, autonomous systems, cybersecurity, governance, and education. In these domains, the ability to interpret AI decisions directly impacts trust, adoption, and accountability. At the same time, evaluation metrics including fidelity, human-centered usability, robustness, and trade-off considerations are vital in ensuring that explanations are not only technically sound but also practically meaningful for end-users. Despite significant progress, open challenges remain, such as the lack of standardization in metrics, the trade-off between accuracy and interpretability, scalability issues, and broader ethical, legal, and social implications (ELSI). These concerns highlight the urgent need for frameworks that balance technical validity with fairness, transparency, and inclusivity. Looking ahead, the path forward for XAI involves a strong focus on human-centered and interactive explanations, hybrid symbolic-statistical approaches, standardized evaluation frameworks, and applications in emerging fields such as federated learning, generative AI, and edge computing. Additionally, policy and regulatory initiatives will play a crucial role in aligning technological advancements with societal values.

In conclusion, explainability is not an auxiliary feature of AI but a cornerstone for trustworthy, sustainable, and ethical AI systems. The continued pursuit of robust XAI methods, supported by interdisciplinary collaboration, will be key to ensuring that artificial intelligence evolves in ways that are transparent, accountable, and beneficial to humanity.

References

[1] M. N. Siddiqui, "AI Revolution: Empowering The Future With Artificial Intelligence," *Pakistan J. Int. Aff.*, vol. 6, no. 3, 2023.
[2] N. L. Rane, M. Paramesha, S. P. Choudhary, and J. Rane, "Artificial intelligence, machine learning, and deep learning for advanced business strategies: a review," *Partners Univers. Int. Innov. J.*, vol. 2, no. 3, pp. 147–171, 2024.
[3] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.
[4] A. Chinnaraju, "Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability," *World J. Adv. Eng. Technol. Sci.*, vol. 14, no. 3, pp. 170–207, 2025.
[5] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (XAI) in healthcare: interpretable models for clinical decision support," *Available SSRN 4637897*, 2023.
[6] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and M. Angulakshmi, "A Literature review on applications of explainable artificial intelligence (XAI)," *IEEE Access*, 2025.
[7] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEe Access*, vol. 11, pp. 78994–79015, 2023.
[8] S. Mirzaei, H. Mao, R. R. O. Al-Nima, and W. L. Woo, "Explainable ai evaluation: A top-down approach for selecting optimal explanations for black box models," *Information*, vol. 15, no. 1, p. 4, 2023.
[9] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, P. E. Ukeoma, and N. F. Dibiaezue, "Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human," *Neural Process. Lett.*, vol. 57, no. 1, p. 16, 2025.
[10] A. Nastoska, B. Jancheska, M. Rizinski, and D. Trajanov, "Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries," *Electronics*, vol. 14, no. 13, p. 2717, 2025.
[11] L. Zhou and M. Sordo, "Expert systems in medicine," in *Artificial intelligence in medicine*, Elsevier, 2021, pp. 75–100.
[12] P. Mishra, D. Henriksen, and J. Dunnigan, "From Symbols to Statistics: The Parallel Histories of Machine and Human Learning," *TechTrends*, pp. 1–8, 2025.
[13] A. Alkhakani, J. Agato, G. Sharma, and P. Adekola, "Harmonizing Data Privacy Laws and AML Requirements: Bridging the Gap Between GDPR, CCPA, and Global Regulations," 2025.
[14] E. Bayamlıoğlu, "The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called 'right to explanation,'" *Regul. Gov.*, vol. 16, no. 4, pp. 1058–1078, 2022.
[15] B. A. Juliussen, "The Right to an Explanation Under the GDPR and the AI Act," in *International Conference on Multimedia Modeling*, Springer, 2025, pp. 184–197.
[16] S. Ma, "Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies," *arXiv Prepr. arXiv2410.21183*, 2024.

[17] V. Pillai, "Enhancing transparency and understanding in AI decision-making processes," *Iconic Res. Eng. Journals*, vol. 8, no. 1, pp. 168–172, 2024.
[18] G. Carloni, A. Berti, and S. Colantonio, "The role of causality in explainable artificial intelligence," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 15, no. 2, p. e70015, 2025.
[19] M. Chromik and A. Butz, "Human-XAI interaction: a review and design principles for explanation user interfaces," in *IFIP Conference on Human-Computer Interaction*, Springer, 2021, pp. 619–640.
[20] U. Farinu, "Fairness, Accountability, and Transparency in AI: Ethical Challenges in Data-Driven Decision-Making," *Available SSRN 5128174*, 2025.
[21] M. A. K. Akhtar, M. Kumar, and A. Nayyar, "Transparency and accountability in explainable AI: Best practices," in *Towards ethical and socially responsible explainable ai: Challenges and opportunities*, Springer, 2024, pp. 127–164.
[22] D. A. Tuan, "Bridging the gap between black box AI and clinical practice: Advancing explainable AI for trust, ethics, and personalized healthcare diagnostics," 2024.
[23] C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach," *Sci. Rep.*, vol. 14, no. 1, p. 25174, 2024.
[24] G. Agrawal, "Accountability, trust, and transparency in AI systems from the perspective of public policy: Elevating ethical standards," in *AI healthcare applications and security, ethical, and legal considerations*, IGI Global, 2024, pp. 148–162.
[25] M. A. K. Akhtar, M. Kumar, and A. Nayyar, "Socially responsible applications of explainable AI," in *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities*, Springer, 2024, pp. 261–350.
[26] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, 2022.
[27] G. Mutahar and T. Miller, "Concept-based explanations using non-negative concept activation vectors and decision tree for cnn models," *arXiv Prepr. arXiv2211.10807*, 2022.
[28] V. Kaffes, D. Sacharidis, and G. Giannopoulos, "Model-agnostic counterfactual explanations of recommendations," in *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, 2021, pp. 280–285.
[29] L. Dib, "Formal definition of interpretability and explainability in XAI," in *Intelligent Systems Conference*, Springer, 2024, pp. 133–151.
[30] D. Bhati, M. Amiruzzaman, Y. Zhao, A. Guercio, and T. Le, "A Survey of Post-Hoc XAI Methods From a Visualization Perspective: Challenges and Opportunities," *IEEE Access*, 2025.
[31] P. Mishra, "Explainability for Linear Models," in *Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks*, Springer, 2021, pp. 35–92.
[32] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Fuzzy rule-based local surrogate models for black-box model explanation," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 6, pp. 2056–2064, 2022.
[33] A. Saini and R. Prasad, "Locally Interpretable Model Agnostic Explanations using Gaussian Processes," *ArXiv. abs/2108.06907*, 2021.
[34] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Auton. Intell. Syst.*, vol. 4, no. 1, p. 2, 2024.
[35] Y. Gao, J. Liu, W. Li, M. Hou, Y. Li, and H. Zhao, "Augmented grad-cam++: super-resolution saliency maps for visual interpretation of deep neural network," *Electronics*, vol. 12, no. 23, p. 4846, 2023.
[36] C. Molnar *et al.*, "Relating the partial dependence plot and permutation feature importance to the data generating process," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 456–479.
[37] K. Eskandar, "Artificial intelligence in healthcare: explore the applications of AI in various medical domains, such as medical imaging, diagnosis, drug discovery, and patient care," *Ser. Med Sci*, vol. 4, pp. 37–53, 2023.
[38] T. Aftab, M. Hussain, M. A. Saeed, A. Yousaf, N. A. Shah, and H. Ahmed, "XAI and disease diagnosis," in *Explainable Artificial Intelligence (XAI) in Healthcare*, CRC Press, 2024, pp. 100–140.
[39] S. B. Koneti, "Regulation, Ethics and Fairness in Artificial Intelligence for Finance: Governance, Explainability and Compliance," *Artif. Intell. Financ. Algorithms, Anal. Autom. Next Financ. Revolut.*, vol. 4, p. 124, 2025.
[40] R. Mohammed, "Artificial intelligence-driven robotics for autonomous vehicle navigation and safety," *NEXG AI Rev. Am.*, vol. 3, no. 1, pp. 21–47, 2022.
[41] K. Malik, M. Sharma, S. Deswal, U. Gupta, D. Agarwal, and Y. O. B. Al Shamsi, "Explainable Artificial Intelligence for Autonomous Vehicles: Concepts, Challenges, and Applications," 2024.
[42] K. S. Alketbi and A. Mehmood, "A Comprehensive Survey of Explainable Artificial Intelligence Techniques for Malicious Insider Threat Detection," *IEEE Access*, 2025.
[43] H. J. Oberhauser, "Bias in Artificial Intelligence: Exploring its Role in Institutional Discrimination and Strategies for Mitigation." Universidade NOVA de Lisboa (Portugal), 2025.
[44] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-assisted decision-making: A systematic review," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 78–91.
[45] M. Miró-Nicolau, A. Jaume-i-Capó, and G. Moyà-Alcover, "A comprehensive study on fidelity metrics for XAI," *Inf. Process. Manag.*, vol. 62, no. 1, p. 103900, 2025.
[46] O. Sankowski and D. Krause, "The human-Centredness metric: early assessment of the quality of human-Centred design activities," *Appl. Sci.*, vol. 13, no. 21, p. 12090, 2023.
[47] M. Esna-Ashari, "Beyond the Black Box: A Review of Quantitative Metrics for Neural Network Interpretability and Their Practical Implications," *Int. J. Sustain. Appl. Sci. Eng.*, vol. 2, no. 1, pp. 1–24, 2025.
[48] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "XAI systems evaluation: A review of human and computer-centred methods," *Appl. Sci.*, vol. 12, no. 19, p. 9423, 2022.
[49] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine learning models," *J. Reliab. Intell. Environ.*, vol. 11, no. 1, p. 1, 2025.
[50] M. Marchiori Manerba, "Fairness Auditing, Explanation and Debiasing in Linguistic Data and Language Models," 2025.
[51] M. K. Mahto, "Explainable artificial intelligence: Fundamentals, approaches, challenges, XAI evaluation, and validation," in *Explainable Artificial Intelligence for Autonomous Vehicles*, CRC Press, 2024, pp. 25–49.
[52] Y. A. Waykar, "Human-AI collaboration in explainable recommender systems: An exploration of user-centric explanations and evaluation frameworks," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 07, pp. 2582–3930, 2023.
[53] L. N. DeLong, R. F. Mir, and J. D. Fleuriot, "Neurosymbolic AI for reasoning over knowledge graphs: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, 2024.
[54] C. Agarwal *et al.*, "Openxai: Towards a transparent evaluation of model explanations," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 15784–15799, 2022.
[55] B. Lund *et al.*, "Standards, frameworks, and legislation for artificial intelligence (AI) transparency," *AI Ethics*, pp. 1–17, 2025.