International Journal of Scientific World, 11 (1) (2025) 13-20

International Journal of Scientific World

Website: www.sciencepubco.com/index.php/IJSW

Research paper

Unveiling explainability in artificial intelligence: a step to-wards transparent AI

Ridwan Boya Marqas [1, 2] *, Saman M. Almufti [2], Rezhna Azad Yusif [2]

[1] Computer Dept., Technical Institute of Shekhan, Shekhan, Duhok, Iraq
[2] Computer Dept., Knowledge University, Erbil, Iraq
**Corresponding author E-mail: pgmr.red@gmail,com*

Abstract

Explainability in artificial intelligence (AI) is an essential factor for building transparent, trustworthy, and ethical systems, particularly in high-stakes domains such as healthcare, finance, justice, and autonomous systems. This study examines the foundations of AI explainability, its critical role in fostering trust, and the current methodologies used to interpret AI models, such as post-hoc techniques, intrinsically interpretable models, and hybrid approaches. Despite these advancements, challenges persist, including trade-offs between accuracy and interpretability, scalability, ethical risks, and transparency gaps. The paper explores emerging trends like causality-based explanations, neuro-symbolic AI, and personalized frameworks, while emphasizing the integration of ethics and the need for automation in explainability. Future directions stress the importance of collaboration among researchers, practitioners, and policymakers to establish industry standards and regulations, ensuring that AI systems align with societal values and expectations.

Keywords: *Explainable AI; Transparency; Post-Hoc Explanations; Causality-Based Explanations; Neuro-Symbolic AI; Ethics In AI; AI Accountability; Trustworthy AI; AI Interpretability; Autonomous Systems.*

1. Introduction

1.1. Background

Artificial Intelligence (AI) has become an integral part of critical decision-making processes across various sectors. In healthcare, AI-powered diagnostic tools assist clinicians in identifying diseases with greater accuracy and efficiency. Similarly, in finance, AI systems help detect fraudulent activities and automate complex trading strategies. In the justice system, AI algorithms are increasingly being used for risk assessment and sentencing. While these advancements bring unprecedented benefits, the growing reliance on AI in these high-stakes areas underscores the need for reliable and transparent systems.

1.2. Problem statement

Despite AI's potential, its lack of transparency poses significant challenges. Many AI systems, especially those based on complex models like deep learning, function as "black boxes," making their decision-making processes difficult to understand. This opacity leads to ethical concerns, such as bias in decision-making, and raises legal implications, particularly when individuals are affected by AI-driven decisions without a clear explanation. This lack of trust and accountability often impedes the broader adoption of AI technologies.

1.3. Motivation

Explainability is crucial in addressing these challenges. By making AI systems interpretable, stakeholders can understand, evaluate, and trust their decisions. Explainability fosters accountability, enabling developers and users to identify and mitigate biases or errors. Moreover, it enhances the adoption of AI by reassuring regulators, end-users, and policymakers of its fairness and reliability.

1.4. Objectives

The primary objective of this paper is to analyze the current state of explainability in AI and underscore its significance in creating transparent systems. It aims to explore the methods, challenges, and advancements in explainability to provide insights into how AI systems can be made more interpretable and trustworthy.

1.5. Paper outline

- Introduction: Provides an overview of the importance of AI explainability and introduces the paper's focus.
- Importance of Explainability: Examines why explainability is essential, particularly in sensitive domains like healthcare, finance, and justice.
- Current Approaches to Explainability: Discusses existing methods for enhancing explainability in AI, including post-hoc explanations and inherently interpretable models.
- Challenges and Limitations: Highlights the technical, ethical, and practical obstacles in achieving explainable AI.
- Future Directions: Explores potential advancements and frameworks for improving explainability while maintaining AI performance.
- Conclusion: Summarizes key findings and emphasizes the role of explainability in fostering trust and transparency in AI systems.

2. Foundations of explainability in AI

2.1. Definition

In artificial intelligence (AI), explainability refers to the capacity of a system to articulate the reasoning behind its decisions or outputs in a manner comprehensible to humans. This transparency is crucial for fostering trust and facilitating informed decision-making (1). Interpretability, while closely related, pertains to the extent to which a human can understand the cause of a decision made by an AI model. It focuses on the clarity of the model's internal mechanisms, whereas explainability encompasses the broader context, including the communication of the model's behavior to users (2).

2.2. Key concepts

2.2.1. Black-box models vs. white-box models

- Black-box models: These are complex AI systems whose internal workings are not transparent or easily understood by humans. Examples include deep neural networks and ensemble methods, which, despite their high performance, lack interpretability (3).
- White-box models: In contrast, white-box models are inherently interpretable, with transparent internal structures that allow users to understand how inputs are transformed into outputs. Examples include decision trees and linear regression models (4).

2.2.2. Levels of explainability

Explainability can be considered at different levels:

- User-level: Explanations tailored for end-users, focusing on how the AI's decisions affect them and providing insights in a non-technical manner (5).
- Developer-level: Detailed explanations aimed at developers and data scientists, offering insights into the model's internal processes, facilitating debugging, and model improvement (6).

2.2.3. Trade-offs: accuracy vs. interpretability

There is often a trade-off between a model's accuracy and its interpretability. Highly accurate models, such as deep learning networks, tend to be less interpretable, making it challenging to understand their decision-making processes. Conversely, simpler models are more interpretable but may not achieve the same level of accuracy. Balancing this trade-off is crucial, especially in domains where transparency is essential (7).

2.3. Importance

Explainability is critical for several reasons:

- Trust: Transparent AI systems enable users to trust the decisions made by the model, especially in high-stakes areas like healthcare and finance (8).
- Compliance: Regulatory frameworks often require that AI decisions be explainable to ensure fairness and accountability (9).
- Debugging: Developers can identify and rectify errors or biases in the model by understanding its decision-making process (10).
- Accountability: Explainable AI allows for the assignment of responsibility for decisions, which is essential in legal and ethical contexts (11).

2.4. Historical context

The concept of explainable AI (XAI) has evolved over time. Initially, AI systems were relatively simple and inherently interpretable. However, as models became more complex, particularly with the advent of deep learning, the need for explainability grew. This led to the development of various techniques aimed at making AI systems more transparent, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) (12). In recent years, there has been a significant push towards integrating explainability into AI systems from the outset, rather than as an afterthought. This shift is driven by the increasing deployment of AI in critical sectors and the corresponding need for transparency and

3. Current approaches to explainable AI

Explainable AI (XAI) encompasses diverse methodologies designed to improve the transparency and interpretability of AI models. These approaches are generally classified into three main categories: post-hoc methods, intrinsically explainable models, and hybrid approaches. Each method has its strengths and limitations, tailored for specific use cases.

3.1. Post-hoc methods

Post-hoc methods are applied to pre-trained models to provide explanations without altering the model's structure or performance. These techniques are particularly useful for understanding complex black-box models like deep neural networks.

3.1.1. Feature importance

- SHAP (SHapley Additive exPlanations): SHAP assigns importance values to each feature based on Shapley values, a concept from game theory. It provides consistent and locally accurate explanations for individual predictions (1).
- LIME (Local Interpretable Model-agnostic Explanations): LIME creates locally interpretable surrogate models around individual predictions, offering a simplified understanding of the decision boundary for specific inputs (2).

3.1.2. Visual explanations

- Saliency Maps: Highlight regions in the input (e.g., an image) that most influence the model's prediction. They are widely used in computer vision tasks to understand what features the model focuses on (3).
- Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM generates class-specific heatmaps to visualize which regions of an image contribute most to a given class prediction. It is particularly effective for convolutional neural networks (4).

3.1.3. Counterfactual explanations

Counterfactual methods provide explanations by identifying minimal changes to the input that would alter the model's output. For example, a counterfactual explanation in a loan approval system might suggest that increasing the applicant's credit score by a certain number would result in approval (5).

3.2. Intrinsically explainable models

Intrinsically explainable models are designed to be transparent and interpretable by nature. These models have simpler structures that make their decision-making processes easily understandable.

3.2.1. Traditional models

- Decision Trees: These models represent decisions and their possible consequences in a tree-like structure. Their straightforward representation makes them ideal for interpretable applications (6).
- Linear Regression: Linear regression models are interpretable because they provide direct coefficients for each feature, indicating their influence on the outcome (7).
- Rule-Based Systems: Rule-based models use if-then-else rules to make predictions, making their decision-making process explicitly clear (8).

3.2.2. Emerging explainable neural networks

Recent research focuses on developing neural networks with inherently explainable structures. Examples include models that incorporate attention mechanisms to highlight important features or networks that use disentangled representations for better interpretability (9).

3.2.3. Hybrid approaches

Hybrid approaches aim to balance interpretability and performance by integrating aspects of both post-hoc explanations and inherently explainable models.

- Surrogate Models: Train interpretable models, like decision trees, to approximate the behavior of complex models, providing insights without compromising accuracy (10).
- Interpretable Layers in Neural Networks: Some hybrid methods modify neural network architectures by adding interpretable layers, such as attention or feature attribution layers, to provide real-time explanations during inference (11).
- Model Compression: Compress a complex model into a simpler one that mimics its behavior while remaining interpretable (12).

Table 1: Comparison of Methods

Approach	Strengths	Weaknesses	Use Cases
Post-Hoc Methods	No need to modify pre-trained models; versatile for black-box models.	Explanations may lack fidelity to the original model.	Image recognition, tabular data analysis.
Intrinsically Explainable	Transparent by design; easy to understand and debug.	May lack the performance of complex models in high-dimensional data.	Healthcare, finance, regulatory environments
Hybrid Approaches	Combines interpretability with high performance; adaptable to various domains.	Can be computationally expensive and complex to implement.	High-stakes applications requiring both performance and transparency.

The landscape of XAI methods continues to evolve, with each approach offering unique advantages. Post-hoc methods are essential for interpreting black-box models, while intrinsically explainable models are suited for applications where transparency is non-negotiable. Hybrid approaches provide a promising path forward, bridging the gap between performance and interpretability to address the needs of critical domains.

3.3. Evaluation of explainability

Evaluating explainability in AI systems is a complex and multifaceted process that requires a combination of quantitative and qualitative metrics. The effectiveness of explanations often depends on the context, the audience, and the purpose of the AI system, making a standardized evaluation challenging.

3.3.1. Metrics for measuring explainability

- Fidelity measures how accurately an explanation reflects the behavior of the underlying model. High fidelity ensures that the explanation faithfully represents the decision-making process without introducing inaccuracies. For instance, surrogate models like LIME and SHAP aim to achieve high fidelity by approximating the original model's predictions in a locally interpretable manner (1).
- Interpretability assesses the simplicity and clarity of an explanation. Explanations are considered interpretable if they are easy for a target audience (e.g., domain experts or laypersons) to understand. Metrics like the number of rules in a rule-based system or the depth of a decision tree are often used as proxies for interpretability (2).
- Completeness measures how well an explanation captures all the factors contributing to a model's decision. Explanations that omit critical information can lead to incorrect interpretations. Completeness can be evaluated by ensuring that the explanation accounts for the most influential features and their interactions (3).
- Actionability assesses whether the explanation provides meaningful insights that enable users to take specific actions. For example, counterfactual explanations are actionable because they suggest what changes would alter the model's output (4).
- Consistency measures whether similar inputs lead to similar explanations. An inconsistent explanation can undermine trust and hinder decision-making (5).

3.3.2. Challenges in evaluation

- Subjectivity in Defining "Understandable" Explanations The notion of "understandable" explanations is subjective and varies among individuals based on their expertise, cognitive ability, and familiarity with AI systems. What is interpretable for a data scientist might be incomprehensible for a layperson (6).
- Lack of Standardized Benchmarks Unlike accuracy or precision, explainability lacks widely accepted benchmarks. Current evaluation methods are often task-specific and inconsistent across studies, making it difficult to compare different techniques or models (7).
- Trade-offs Between Metrics Improving one metric (e.g., fidelity) can sometimes reduce another (e.g., interpretability). Balancing these trade-offs requires careful consideration of the application context (8).
- Dynamic Contexts Explanations that are effective in one scenario may fail in another. For example, explanations in high-stakes domains like healthcare must be more rigorous than those in recommendation systems. Evaluating such dynamic requirements remains a significant challenge (9).

3.3.3. User-centric evaluation

User-centric evaluation focuses on how well explanations meet the needs of their intended audience. This approach emphasizes the utility and comprehensibility of explanations in real-world contexts.

- Human-Centric Metrics
- Comprehension: Evaluate whether users can accurately describe the model's behavior after reviewing an explanation. Studies often use surveys or interviews to assess this aspect (10).
- Trust: Measure whether explanations increase users' confidence in the model's decisions. Trust is often gauged through experiments that assess user reliance on AI systems in decision-making tasks (11).
- Satisfaction: Assess user satisfaction with the provided explanations. This metric is typically collected through user feedback and surveys (12).
- Actionability is tested by evaluating whether users can make better decisions after understanding the explanation. For example, in credit scoring, an explanation that highlights actionable factors (e.g., improving credit history) can be assessed by its effectiveness in guiding user decisions (13).
- Simulatability evaluates whether users can simulate the AI model's predictions after seeing an explanation. This metric tests if the explanation provides sufficient clarity for users to predict the model's behavior in new scenarios (14).

Evaluating explainability requires a multidimensional approach that combines objective metrics like fidelity and completeness with subjective, human-centric assessments such as comprehension and trust. Despite advancements, challenges like subjectivity, lack of benchmarks, and contextual variability remain. A standardized evaluation framework that integrates quantitative and qualitative measures is essential to advance the field and build trust in AI systems.

3.4. Real-world applications of explainability

Explainability in artificial intelligence (AI) has emerged as a crucial factor for its deployment in critical domains. These applications require models to provide interpretable and actionable insights, ensuring transparency, trust, and compliance with legal and ethical standards.

3.4.1. Examples of explainability in critical domains

1) Healthcare

- Application: Explainability plays a significant role in interpreting diagnostic decisions in medical imaging. For instance, saliency maps and Grad-CAM are used in deep learning models to highlight regions in medical scans (e.g., X-rays, MRIs) that contribute most to a diagnosis. This helps clinicians verify AI recommendations and reduces the risk of misdiagnosis (1).
- Example: AI models for predicting cardiovascular disease risk often rely on explainable methods like SHAP or LIME to identify key contributing factors, such as age, cholesterol levels, or blood pressure (2).
- Challenges: Healthcare systems face domain-specific issues such as patient data sensitivity, regulatory compliance (e.g., HIPAA), and the life-critical nature of decisions, where inaccuracies can lead to severe consequences.

2) Finance

- Application: In finance, explainability is essential for credit scoring models, fraud detection systems, and algorithmic trading. For instance, LIME can be used to explain why a loan application was rejected, based on factors like credit history or debt-to-income ratio (3).

- Example: Fraud detection systems employ counterfactual explanations to identify unusual patterns in transaction data and provide actionable insights for financial analysts (4).
- Challenges: Financial AI must comply with strict regulations, such as the European Union's General Data Protection Regulation (GDPR), which mandates the right to an explanation for automated decisions. Moreover, the complexity of financial systems often leads to trade-offs between interpretability and performance.

3) Justice

- Application: In criminal justice, explainable AI is used to ensure fairness and transparency in risk assessment tools for bail decisions or sentencing recommendations. Models like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) have come under scrutiny for potential biases, highlighting the need for interpretable algorithms (5).
- Example: Counterfactual explanations can be used to ensure fairness by identifying biased inputs, such as race or gender, and suggesting neutral alternatives (6).
- Challenges: The justice system's reliance on explainable AI is complicated by legal and ethical considerations, including the potential for reinforcing societal biases, lack of standardized fairness metrics, and the difficulty of integrating contextual nuances into AI systems.

4) Autonomous systems

- Application: Explainability in autonomous systems, such as self-driving cars and drones, is critical for justifying actions and ensuring safety. For instance, visual explanations like Grad-CAM can help developers understand why a self-driving car recognized a pedestrian or misinterpreted an object (7).
- Example: In aviation, explainable AI systems in drones can provide mission-critical feedback on decision-making processes, such as flight path adjustments or obstacle avoidance (8).
- Challenges: Autonomous systems face challenges related to real-time decision-making, where delays in generating explanations can impact safety. Additionally, ensuring explainability in highly dynamic environments with multiple interacting components is complex.

3.5. Challenges in domain-specific applications

1) Legal regulations

Each domain operates under unique regulatory frameworks. For example, healthcare models must comply with privacy laws like HIPAA, while financial models are subject to GDPR and other accountability standards. These regulations often require AI systems to provide clear and actionable explanations (9).

2) Data sensitivity

Sensitive data, such as medical records or financial transactions, imposes strict constraints on how explanations can be generated and shared. Balancing transparency with privacy remains a key challenge (10).

3) Complexity of models and decisions

In high-stakes applications, the complexity of decisions often demands highly accurate models, which are typically less interpretable. Achieving a balance between accuracy and interpretability is particularly challenging in dynamic environments like autonomous systems (11).

4) Bias and fairness

Domain-specific biases can influence model predictions and explanations. For instance, historical biases in justice system data can lead to unfair sentencing recommendations. Mitigating these biases while maintaining transparency is an ongoing challenge (12).

The integration of explainable AI in critical domains like healthcare, finance, justice, and autonomous systems is essential for fostering trust, accountability, and compliance. However, achieving explainability in these domains requires addressing unique challenges, including legal constraints, data sensitivity, and the inherent trade-offs between interpretability and accuracy. Moving forward, domain-specific frameworks and standardized evaluation metrics will be crucial for ensuring the ethical and effective deployment of explainable AI.

4. Challenges and limitations of current explainability techniques

Despite significant advancements in explainable AI (XAI), many challenges and limitations persist. These challenges hinder the broader adoption of AI in critical domains and raise questions about the reliability and ethical implications of current explainability techniques.

4.1. Trade-offs: accuracy vs. interpretability

One of the most significant challenges in explainability is the trade-off between model accuracy and interpretability.

- Complex Models and Black-Box Nature: Highly accurate models, such as deep neural networks, often operate as "black boxes" with opaque decision-making processes. While these models outperform simpler, interpretable ones in tasks like image recognition and natural language processing, their complexity limits their interpretability (1).
- Simple Models and Loss of Accuracy: Intrinsically interpretable models like decision trees and linear regression are easier to understand but may underperform when handling high-dimensional or unstructured data. This trade-off forces developers to choose between performance and transparency, particularly in high-stakes applications like healthcare or finance (2).
- Efforts to Balance Both: Hybrid approaches, such as interpretable layers or surrogate models, aim to balance accuracy and interpretability but often come at the cost of computational efficiency or partial fidelity (3).

4.2. Scalability

Scaling explainability techniques to complex AI systems poses a significant challenge, particularly in models with millions or billions of parameters.

- High-Dimensional Data: As models grow in size and complexity, generating explanations that are both meaningful and concise becomes increasingly difficult. For example, explaining a single prediction in a deep learning model may require analyzing thousands of interdependent parameters (4).

- Time and Computational Costs: Techniques like SHAP or LIME are computationally intensive, especially when applied to large datasets or deep models. This limits their feasibility for real-time applications or for analyzing multiple predictions simultaneously (5).
- Challenges in Dynamic Systems: Real-time systems, such as autonomous vehicles or financial trading algorithms, require instant explanations. Current methods struggle to provide scalable, actionable insights within these time constraints (6).

4.3. Ethical concerns

- Explainability techniques can introduce ethical risks, including selective explanations, manipulation, and misuse.
- Selective Explanations: Post-hoc methods like LIME or SHAP can be manipulated to highlight specific aspects of a model's behavior while hiding others. This "cherry-picking" of explanations can mislead stakeholders and undermine trust in AI systems (7).
- Bias Amplification: Explainable models are not immune to biases present in the data or the model itself. Misinterpretation of explanations can lead to biased decisions, especially in sensitive domains like criminal justice or hiring (8).
- Misuse by Malicious Actors: Explainability techniques can inadvertently expose model vulnerabilities. For instance, attackers can use explanations to craft adversarial examples that exploit weaknesses in the model (9).

4.4. Transparency gaps

- While explainability techniques provide insights into AI behavior, they often fall short in several critical areas:
- Global Explanations: Most post-hoc methods focus on local explanations, which describe individual predictions rather than providing a global understanding of the model's behavior. This leaves significant gaps in understanding the overall logic and limitations of the model (10).
- Causal Insights: Current techniques often lack the ability to infer causal relationships, focusing instead on correlations. Without causal explanations, it is difficult to identify the root causes of a model's decisions, which is critical for interventions in domains like healthcare and policy (11).
- Human-Centric Explanations: Many methods are designed from a technical perspective and fail to prioritize user-centric explanations that are actionable and meaningful for non-expert users (12).
- Incomplete Feature Representations: Techniques like SHAP and LIME rely heavily on feature importance, which may oversimplify complex interactions between variables or fail to capture nuanced dependencies (13).

Current explainability techniques, while valuable, face significant challenges that limit their effectiveness and trustworthiness in critical domains. Addressing trade-offs between accuracy and interpretability, enhancing scalability, mitigating ethical concerns, and closing transparency gaps are essential steps toward advancing XAI. Future research must focus on developing standardized frameworks, improving causal inference, and prioritizing human-centric design to ensure that explainability techniques align with the needs of diverse stakeholders.

5. Towards transparent ai: the future of explainability

The future of explainability in artificial intelligence (AI) is pivotal for achieving transparent, ethical, and user-centric AI systems. As AI becomes increasingly integrated into critical decision-making processes, new approaches and frameworks are emerging to address current limitations and enhance the interpretability of these systems.

5.1. Emerging trends

1) Causality-based explanations
- Overview: Traditional explainability techniques often rely on correlation rather than causation. Causality-based explanations aim to uncover the cause-and-effect relationships driving AI decisions, providing deeper and more actionable insights (1).
- Impact: By focusing on causal mechanisms, these approaches enable more robust decision-making, particularly in domains like healthcare and policy, where understanding root causes is critical (2).

2) Neuro-symbolic AI
- Overview: Neuro-symbolic AI combines the learning capabilities of neural networks with the logical reasoning of symbolic systems. This hybrid approach enhances explainability by integrating interpretable rules with powerful pattern recognition (3).
- Applications: For instance, neuro-symbolic systems can explain predictions in natural language processing tasks by mapping neural outputs to symbolic knowledge representations (4).

3) Personalized explainability for different user groups
- Overview: The same explanation may not be effective for all users. Personalized explainability tailors explanations to meet the needs of diverse audiences, such as domain experts, regulators, or laypersons (5).
- Approach: Techniques include user profiling and adaptive explanation systems that adjust the level of detail, technical complexity, and presentation format based on the user's preferences and expertise (6).

5.2 Integration with ethics

1) Fairness and accountability
- Alignment: Future explainability methods must ensure that explanations promote fairness by revealing and mitigating biases in AI models (7).
- Example: Counterfactual explanations can highlight unfair treatment by demonstrating how changing sensitive attributes (e.g., gender or ethnicity) might alter outcomes (8).

2) Privacy
- Balancing Transparency and Privacy: Explainability must respect data privacy, especially in domains like healthcare and finance. Techniques such as federated learning and differential privacy can help achieve this balance (9).
- Challenge: Generating explanations without exposing sensitive data remains an ongoing area of research (10).

5.3. Automation in explainability

1) Dynamic explanation generation
- Overview: AI systems are evolving to dynamically generate explanations for their decisions, adapting to new scenarios and user queries in real time (11).
- Advancements: Methods such as reinforcement learning and meta-learning are being explored to enable AI systems to learn how to explain their outputs effectively (12).
- Benefits: Dynamic explanations improve user trust and understanding, particularly in complex or evolving applications like autonomous vehicles or real-time fraud detection (13).

2) Self-explaining AI models
- Concept: Self-explaining models are designed to inherently justify their predictions without the need for external interpretability methods. Examples include neural networks with interpretable intermediate layers and models that embed explainability as a core design principle (14).
- Impact: These systems provide real-time, actionable insights while maintaining high performance (15, 16,17).

5.4. Standards and regulations

1) Industry standards
- Need for Standardization: The absence of standardized frameworks for evaluating and implementing explainability leads to inconsistencies across AI systems. Developing industry-wide standards is critical for ensuring uniformity and reliability (16).
- Example Initiatives: Organizations like IEEE and ISO are working on guidelines for AI ethics and transparency, including metrics for explainability (17).

2) Regulatory compliance
- Call for Guidelines: Governments and regulatory bodies are increasingly emphasizing the importance of explainability in AI systems. For instance, the European Union's General Data Protection Regulation (GDPR) includes a "right to explanation" for automated decisions (18,20).
- Challenges: Striking a balance between explainability requirements and technological feasibility remains a challenge, especially for complex models (19,21).

The future of explainability lies in developing transparent, adaptable, and ethical AI systems. Emerging trends such as causality-based explanations, neuro-symbolic AI, and personalized explainability are paving the way for more meaningful insights. Integrating explainability with ethical principles like fairness, accountability, and privacy will ensure that AI systems align with societal values. Additionally, automation in explainability and the establishment of industry standards and regulations will play crucial roles in fostering trust and widespread adoption. By addressing these advancements, the field moves closer to achieving transparent AI that benefits all stakeholders.

6. Conclusion

Explainability in artificial intelligence (AI) is not merely a technical challenge but a foundational requirement for building trust, fostering accountability, and ensuring the ethical deployment of AI systems. As AI increasingly permeates critical domains like healthcare, finance, justice, and autonomous systems, the need for clear, interpretable, and actionable insights has become paramount.

This discussion has highlighted the range of approaches to explainability, from post-hoc methods like SHAP and Grad-CAM to inherently interpretable models such as decision trees and emerging hybrid frameworks that aim to balance accuracy with transparency. Despite these advancements, significant challenges persist, including the trade-off between model performance and interpretability, scalability issues, ethical concerns, and gaps in transparency such as the lack of global explanations and causal insights.

Future directions in explainability focus on causality-based explanations, neuro-symbolic AI, personalized user-centric frameworks, and automation in generating dynamic, self-explaining models. Equally important is the alignment of explainability efforts with ethical principles like fairness, accountability, and privacy, alongside the development of industry standards and regulatory compliance mechanisms.

Achieving transparent AI requires collaboration across disciplines. Researchers must continue to innovate and address technical limitations; practitioners must prioritize user-centric designs in real-world applications; and policymakers must establish clear guidelines to ensure accountability and fairness. By uniting these efforts, we can advance the field of explainable AI and ensure that its transformative potential is realized in a manner that aligns with societal values and expectations.

References

[1] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[2] Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778.

[4] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116.

[5] S. M. Abdulrahman, R. R. Asaad, H. B. Ahmad, A. Alaa Hani, S. R. M. Zeebaree, and A. B. Sallow, "Machine Learning in Nonlinear Material Physics," Journal of Soft Computing and Data Mining, vol. 5, no. 1, Jun. 2024, https://doi.org/10.30880/jscdm.2024.05.01.010.

[6] A. B. Sallow, R. R. Asaad, H. B. Ahmad, S. Mohammed Abdulrahman, A. A. Hani, and S. R. M. Zeebaree, "Machine Learning Skills To K–12," Journal of Soft Computing and Data Mining, vol. 5, no. 1, Jun. 2024, https://doi.org/10.30880/jscdm.2024.05.01.011.

[7] S. M. Almufti et al., "INTELLIGENT HOME IOT DEVICES: AN EXPLORATION OF MACHINE LEARNING-BASED NETWORKED TRAFFIC INVESTIGATION," Jurnal Ilmiah Ilmu Terapan Universitas Jambi, vol. 8, no. 1, pp. 1–10, May 2024, doi: 10.22437/jiituj.v8i1.32767. https://doi.org/10.22437/jiituj.v8i1.32767.

[8] S. M. Almufti and S. R. M. Zeebaree, "Leveraging Distributed Systems for Fault-Tolerant Cloud Computing: A Review of Strategies and Frameworks," Academic Journal of Nawroz University, vol. 13, no. 2, pp. 9–29, May 2024, https://doi.org/10.25007/ajnu.v13n2a2012.

[9] H. B. Ahmad, R. R. Asaad, S. M. Almufti, A. A. Hani, A. B. Sallow, and S. R. M. Zeebaree, "SMART HOME ENERGY SAVING WITH BIG DATA AND MACHINE LEARNING," Jurnal Ilmiah Ilmu Terapan Universitas Jambi, vol. 8, no. 1, pp. 11–20, May 2024, https://doi.org/10.22437/jiituj.v8i1.32598.

[10] T. Thirugnanam et al., "PIRAP: Medical Cancer Rehabilitation Healthcare Center Data Maintenance Based on IoT-Based Deep Federated Collaborative Learning," Int J Coop Inf Syst, vol. 33, no. 01, Mar. 2024, https://doi.org/10.1142/S0218843023500053.
[11] R. Boya Marqas, S. M. Almufti, and R. Rajab Asaad, "FIREBASE EFFICIENCY IN CSV DATA EXCHANGE THROUGH PHP-BASED WEBSITES," Academic Journal of Nawroz University, vol. 11, no. 3, pp. 410–414, Aug. 2022, https://doi.org/10.25007/ajnu.v11n3a1480.
[12] S. M. Almufti, R. B. Marqas, Z. A. Nayef, and T. S. Mohamed, "Real Time Face-mask Detection with Arduino to Prevent COVID-19 Spreading," Qubahan Academic Journal, vol. 1, no. 2, pp. 39–46, Apr. 2021, https://doi.org/10.48161/qaj.v1n2a47.
[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (pp. 4765-4774).
[14] Zhang, J., & Harman, M. (2021). Interpretable machine learning: A survey. arXiv preprint arXiv:2103.11251.
[15] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887. https://doi.org/10.2139/ssrn.3063289.
[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
[17] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314.
[18] Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850.
[19] Shapley, L. S. (1953). A value for n-person games. In Contributions to the Theory of Games (pp. 307-317). Princeton University Press. https://doi.org/10.1515/9781400881970-018.
[20] Marqas, R. B., Mousa, A., Özyurt, F., & Salih, R. (2023). A machine learning model for the prediction of heart attack risk in high-risk patients utilizing real-world data. Academic Journal of Nawroz University, 12(4), 286-301. https://doi.org/10.25007/ajnu.v12n4a1974.
[21] Abdalla Mohammed Abubakr, A., Khan, F., Alhag Ali Mohammed, A., Abdelbagi Abdalla, Y., Abd Alla Mohammed, A., & Ahmad, Z. (2024). Impact of AI applications on corporate financial reporting quality: Evidence from UAE corporations. Qubahan Academic Journal, 4(3), 782–792. https://doi.org/10.48161/qaj.v4n3a860.