Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 08| 2025

Research Article

Received: 06-07-2025| Accepted: 26-07-2025 | Published: 14-08-2025

Explainable AI: Building Trust and Transparency in Machine Learning Models

Amit Taneja

UMB bank, USA

Abstract: Artificial intelligence systems increasingly influence critical decisions across healthcare, finance, and judicial domains, yet their complex architectures often operate as opaque black boxes. This creates significant barriers to user acceptance and regulatory compliance in high-stakes applications where understanding decision rationale proves essential. Explainable AI emerges as a fundamental requirement for building trustworthy systems that balance predictive performance with interpretability. Various techniques address interpretability challenges, including model-agnostic methods like LIME and SHAP, intrinsically interpretable architectures, and post-hoc explanation frameworks. However, significant technical challenges persist, including the accuracy-interpretability trade-off, computational overhead, and ensuring explanation fidelity. Human factors play crucial roles in determining explanation effectiveness, with user cognitive limitations and domain expertise influencing trust calibration. Successful implementation requires addressing both technical capabilities and human-centered design principles. The regulatory landscape increasingly demands algorithmic accountability, making explainability not merely beneficial but mandatory for many applications. Transparent AI systems demonstrate potential for enhanced user adoption, improved decision quality, and ethical compliance. Future developments must prioritize standardized evaluation metrics, scalable explanation methods, and domain-specific interpretability frameworks to realize the full potential of trustworthy artificial intelligence in society.

Keywords: Explainable artificial intelligence, interpretable machine learning, algorithmic transparency, user trust, model interpretability.

INTRODUCTION

Defining Explainable AI and Its Distinction from Interpretable Machine Learning

The advent of sophisticated machine learning algorithms has ushered in an era where artificial intelligence systems demonstrate remarkable predictive capabilities across diverse domains. Explainable Artificial Intelligence (XAI) encompasses techniques and methodologies designed to make artificial intelligence systems more transparent, interpretable, and trustworthy to human users (Ortigossa, E. S. *et al*., 2024). While often used interchangeably, explainable AI and interpretable machine learning represent distinct yet complementary concepts. Interpretable machine learning focuses on creating models whose internal mechanisms can be understood directly, such as linear regression or decision trees. In contrast, explainable AI encompasses a broader scope, including post-hoc explanation methods that can be applied to any model, regardless of its inherent interpretability.

The Growing Complexity of AI Systems and the Corresponding Opacity Problem

The growing complexity of contemporary AI systems exacerbates the opacity problem significantly. Deep neural networks, ensemble methods, and other sophisticated architectures achieve unprecedented accuracy but sacrifice interpretability in the process. This black-box nature creates substantial barriers to understanding, validation, and trust, particularly problematic when these systems influence consequential decisions affecting human lives and societal outcomes. The complexity-interpretability trade-off represents one of the most pressing challenges in modern machine learning, where achieving higher predictive performance often requires sacrificing human comprehensibility.

Critical Domains Where Decision Transparency Is Essential

Decision transparency becomes particularly crucial in domains where AI systems directly impact human welfare and rights. Healthcare applications, including diagnostic systems and treatment recommendation algorithms, demand clear explanations to ensure patient safety and enable physician validation of AI-generated insights. Financial institutions deploying credit scoring algorithms and automated trading systems require interpretability for regulatory compliance and risk management. Criminal justice systems utilizing risk assessment tools for sentencing and parole decisions necessitate transparent explanations to uphold principles of fairness and due process. Autonomous systems, including self-driving vehicles and robotic applications, require explainable decision-making mechanisms to ensure safety and accountability in dynamic environments (Rendón, L. G. 2022).

RESEARCH OBJECTIVES AND ARTICLE STRUCTURE

The primary objective of this work is to provide a comprehensive examination of explainable AI

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

*Corresponding Author: Amit Taneja

DOI- https://doi.org/10.5281/zenodo.16872119

methodologies, challenges, and applications across critical domains. The analysis encompasses technical approaches to interpretability, human factors influencing explanation effectiveness, and trust mechanisms that facilitate user acceptance of transparent AI systems. This article is structured to provide systematic coverage of explainable AI from foundational concepts to practical applications, examining the imperative for AI interpretability in high-stakes applications, exploring taxonomies and methodologies for explainable AI, analyzing technical challenges and limitations, and investigating trust mechanisms and user acceptance factors.

THE IMPERATIVE FOR AI INTERPRETABILITY IN HIGH-STAKES APPLICATIONS

Medical Diagnosis and Treatment Recommendation Systems

Healthcare represents one of the most critical domains demanding AI interpretability, where algorithmic decisions directly impact patient outcomes and physician trust. Medical diagnosis systems powered by machine learning algorithms must provide clear explanations for their diagnostic recommendations to enable healthcare professionals to validate, understand, and confidently act upon AI-generated insights (Sushma, M. *et al*., 2021). Treatment recommendation systems require transparency to ensure that proposed interventions align with clinical reasoning and established medical protocols. The life-or-death nature of medical decisions necessitates that AI systems not only achieve high accuracy but also demonstrate their reasoning process in ways that medical professionals can comprehend and verify against their clinical expertise.

Financial Credit Scoring and Algorithmic Trading Decisions

The financial sector increasingly relies on sophisticated AI algorithms for credit scoring, loan approval processes, and high-frequency algorithmic trading operations. Credit scoring systems must provide explanations for approval or denial decisions to comply with fair lending regulations and enable applicants to understand factors influencing their creditworthiness. Algorithmic trading systems require interpretability to manage risk exposure, comply with regulatory oversight, and maintain investor confidence. The potential for significant financial losses and market disruption demands that these systems operate with sufficient transparency to enable human oversight and intervention when necessary.

Legal and Judicial Decision Support Systems

Criminal justice systems employ AI-powered risk assessment tools for sentencing, parole decisions, and pretrial release determinations, creating an urgent need for algorithmic transparency. These systems must provide clear explanations for their risk predictions to uphold principles of due process and enable judicial review of algorithmic recommendations. The fundamental requirement for fairness and accountability in legal proceedings demands that AI systems operating in this domain demonstrate their decision-making process transparently. Legal professionals and defendants have the right to understand and challenge the basis of algorithmic assessments that influence judicial outcomes.

Autonomous Vehicle Safety-Critical Operations

Autonomous vehicles represent a paradigmatic example of safety-critical AI systems where interpretability becomes essential for public acceptance and regulatory approval (Ma, Y. *et al*., 2020). These systems must explain their decision-making processes during critical situations such as emergency braking, lane changes, and collision avoidance scenarios. The ability to understand why an autonomous vehicle made specific decisions becomes crucial for accident investigation, liability determination, and continuous system improvement. Public trust in autonomous vehicles depends significantly on the transparency of their decision-making algorithms and the ability to provide post-incident explanations.

Regulatory Compliance Requirements

Contemporary regulatory frameworks worldwide increasingly mandate algorithmic transparency and accountability across various domains. The General Data Protection Regulation establishes the right to explanation for automated decision-making systems affecting individuals. Financial regulatory bodies require transparency in algorithmic trading and credit decision systems to prevent market manipulation and ensure fair lending practices. The Food and Drug Administration demands explainability in AI-powered medical devices to ensure patient safety and enable clinical validation. These regulatory requirements create legal imperatives for AI interpretability that extend beyond technical considerations to encompass compliance and governance obligations.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Ethical Considerations and Algorithmic Bias Detection

AI interpretability serves as a fundamental mechanism for identifying and mitigating algorithmic bias across all high-stakes applications. Transparent AI systems enable stakeholders to examine decision patterns for discriminatory behavior against protected groups and ensure equitable treatment across diverse populations. The ability to understand AI decision-making processes facilitates the detection of unintended biases that may emerge from training data or algorithmic design choices. Ethical AI deployment requires not only fair outcomes but also transparent processes that can be audited and validated for compliance with ethical principles and societal values.

Table 1: Critical Domains Requiring AI Interpretability and Their Specific Requirements [(Rendón, L. G. 2022; Ma, Y. *et al*., 2020; Sushma, M. *et al*., 2021)

Domain	Application Examples	Interpretability Requirements	Regulatory Framework	Key Stakeholders
Healthcare	Diagnostic systems, Treatment recommendations	Clinical validation, Safety verification	FDA approval processes	Physicians, Patients, Regulators
Finance	Credit scoring, Algorithmic trading	Fair lending compliance, Risk transparency	GDPR, Financial regulations	Loan officers, Investors, Regulators
Criminal Justice	Risk assessment, Sentencing support	Due process, Bias detection	Constitutional requirements	Judges, Defendants, Legal advocates
Autonomous Vehicles	Safety-critical decisions, Emergency responses	Accident investigation, Liability determination	Transportation safety standards	Engineers, Regulators, Public
General Applications	Automated decision-making	Right to explanation	GDPR Article 22	Data subjects, Controllers

TAXONOMIES AND METHODOLOGIES FOR EXPLAINABLE AI

Model-Agnostic Explanation Techniques

Model-agnostic explanation techniques provide interpretability solutions that can be applied to any machine learning model regardless of its internal architecture or complexity. Local Interpretable Model-agnostic Explanations (LIME) generates explanations by approximating the behavior of complex models with simpler, interpretable models in the local neighborhood of individual predictions (Rao, S. *et al*., 2022). SHapley Additive exPlanations (SHAP) employs game theory principles to assign important values to each feature, providing consistent and theoretically grounded explanations for model predictions. Permutation importance measures feature relevance by evaluating the decrease in model performance when individual features are randomly shuffled, offering a straightforward approach to understanding feature contributions across different model types.

Intrinsically Interpretable Models

Intrinsically interpretable models possess inherent transparency in their decision-making processes, enabling direct human comprehension without additional explanation techniques. Linear models, including linear regression and logistic regression, provide straightforward interpretability through coefficient weights that directly indicate feature importance and direction of influence. Decision trees offer intuitive rule-based explanations through their hierarchical structure, allowing users to trace decision paths from root to leaf nodes. Rule-based systems generate explicit if-then statements that capture decision logic in human-readable formats, facilitating direct understanding of model reasoning processes.

Post-Hoc Explanation Methods

Post-hoc explanation methods generate interpretations after model training and deployment, providing insights into pre-existing models without modifying their internal structure. Gradient-based attribution methods compute feature importance by analyzing gradients of model outputs concerning input features, revealing which aspects of the input most strongly influence predictions. Counterfactual explanations identify minimal changes to input features that would alter model predictions, providing actionable insights into decision boundaries and feature sensitivity.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

These approaches enable interpretability for complex models while preserving their predictive performance and architectural integrity.

Global Versus Local Interpretability Approaches

Global interpretability approaches aim to understand overall model behavior across the entire dataset, providing insights into general patterns and feature relationships that govern model decisions (Islam, M. R. *et al*., 2021). These methods reveal systematic biases, feature interactions, and decision strategies that characterize model behavior at a population level. Local interpretability approaches focus on explaining individual predictions, providing specific insights into why a model made particular decisions for given instances. The distinction between global and local interpretability reflects different stakeholder needs, with global explanations serving model developers and auditors, while local explanations benefit end-users seeking to understand specific predictions.

Visual Explanation Techniques and Saliency Mapping

Visual explanation techniques leverage human visual processing capabilities to communicate model reasoning through graphical representations and interactive visualizations. Saliency mapping highlights regions of input data that most strongly influence model predictions, particularly valuable for image classification and computer vision applications. Heat maps and attention visualizations reveal spatial patterns of feature importance, enabling intuitive understanding of model focus areas. These visual approaches bridge the gap between complex mathematical computations and human cognitive capabilities, making AI explanations more accessible to non-technical stakeholders.

Natural Language Explanations and Conversational AI Transparency

Natural language explanations translate model reasoning into human-readable text, providing intuitive and accessible interpretations of AI decisions. These systems generate narrative explanations that describe feature contributions, decision rationale, and confidence levels in conversational formats. Conversational AI transparency extends this concept by enabling interactive dialogue between users and explanation systems, allowing stakeholders to ask follow-up questions and request clarifications. This approach democratizes AI interpretability by removing technical barriers and enabling natural communication about model behavior and reasoning processes.

Table 2: Taxonomy of Explainable AI Methodologies and Their Characteristics (Ortigossa, E. S. *et al*., 2024; Rao, S. *et al*., 2022; Islam, M. R. *et al*., 2021; Griffiths, T. L. 2020)

Category	Technique	Scope	Model Compatibility	Output Format	Computational Cost
Model-Agnostic	LIME	Local	Universal	Feature importance scores	High
Model-Agnostic	SHAP	Local/Global	Universal	Shapley values	High
Model-Agnostic	Permutation Importance	Global	Universal	Feature rankings	Medium
Intrinsically Interpretable	Linear Models	Global	Self-contained	Coefficient weights	Low
Intrinsically Interpretable	Decision Trees	Global	Self-contained	Decision rules	Low
Post-hoc	Gradient Attribution	Local	Neural networks	Input gradients	Medium
Post-hoc	Counterfactual	Local	Universal	Alternative scenarios	High
Visual	Saliency Mapping	Local	Vision models	Heat maps	Medium

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

TECHNICAL CHALLENGES AND LIMITATIONS IN AI EXPLAINABILITY

The Accuracy-Interpretability Trade-off in Complex Models

The fundamental tension between model accuracy and interpretability represents one of the most persistent challenges in explainable AI. Complex models such as deep neural networks and ensemble methods achieve superior predictive performance through intricate architectures that sacrifice human comprehensibility. This trade-off forces practitioners to choose between deploying highly accurate but opaque models or accepting reduced performance for the sake of interpretability. The challenge becomes particularly acute in domains where both high accuracy and transparency are critical requirements, necessitating innovative approaches that can bridge this gap without compromising either objective.

Computational Overhead and Scalability Issues

Explanation generation often imposes significant computational burdens that can limit the practical deployment of explainable AI systems. Model-agnostic techniques require multiple model evaluations to generate explanations, creating substantial overhead for large-scale applications. The computational cost of explanation methods scales poorly with dataset size and model complexity, potentially making real-time explanation generation infeasible for time-sensitive applications. These scalability limitations constrain the adoption of explainable AI in production environments where computational resources are limited and response time requirements are stringent.

Explanation of Fidelity and Faithfulness to Model Behavior

Ensuring that explanations accurately reflect actual model behavior presents a critical challenge for XAI systems. Explanation methods may generate plausible but misleading interpretations that do not correspond to the model's true decision-making process. The fidelity problem becomes particularly complex when explanation techniques approximate model behavior through simplified representations or local approximations. Distinguishing between explanations that appear reasonable to humans and those that faithfully represent model reasoning requires sophisticated evaluation frameworks and validation methodologies.

Human Cognitive Limitations in Processing Explanations

Human cognitive constraints significantly influence the effectiveness of AI explanations, creating barriers to the successful interpretation and utilization of explainable AI systems (Griffiths, T. L. 2020). Users possess limited working memory capacity, attention span, and ability to process complex multidimensional information simultaneously. Cognitive biases and heuristics affect how individuals interpret and act upon explanations, potentially leading to overconfidence or misunderstanding of model capabilities. The design of effective explanations must account for these human limitations by presenting information in cognitively accessible formats that align with natural reasoning processes.

Adversarial Explanations and Manipulation Vulnerabilities

Adversarial attacks targeting explanation systems pose emerging security threats that can undermine trust in explainable AI. Malicious actors may manipulate input features or model parameters to generate deceptive explanations that obscure true model behavior or create false impressions of fairness and reliability (Jeanneret, G. *et al*., 2023). These adversarial explanations can be particularly dangerous in high-stakes applications where stakeholders rely on explanations for critical decisions. The vulnerability of explanation methods to manipulation highlights the need for robust evaluation frameworks and defensive mechanisms to ensure explanation integrity.

Standardization and Evaluation Metrics for Explanation Quality

The absence of standardized evaluation metrics and benchmarks for explanation quality impedes progress in explainable AI development and deployment. Different explanation methods employ varying approaches to measuring explanation effectiveness, making comparative evaluation difficult. The subjective nature of explanation quality assessment complicates the development of objective metrics that can capture human preferences and domain-specific requirements. Establishing consensus on evaluation criteria requires interdisciplinary collaboration between technical experts, domain specialists, and end-users to define meaningful measures of explanation utility and effectiveness.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Table 3: Technical Challenges in AI Explainability and Their Impact (Ortigossa, E. S. *et al*., 2024; Griffiths, T. L. 2020; Jeanneret, G. *et al*., 2023)

Challenge Category	Specific Issues	Impact on Deployment	Affected Stakeholders	Mitigation Strategies
Accuracy-Interpretability Trade-off	Performance degradation with transparency	Reduced adoption in performance-critical applications	ML Engineers, End-users	Hybrid approaches, Progressive disclosure
Computational Overhead	High explanation generation costs	Limited real-time applications	System administrators, Users	Approximation methods, Caching
Explanation Fidelity	Misleading or inaccurate explanations	False confidence, Poor decisions	Domain experts, Decision-makers	Validation frameworks, Fidelity metrics
Human Cognitive Limitations	Information overload, Cognitive biases	Misinterpretation of explanations	All user groups	Cognitive-aware design, User training
Adversarial Vulnerabilities	Manipulated explanations	Security risks, Trust erosion	Security teams, Auditors	Robust explanation methods, Detection systems
Standardization Gap	Inconsistent evaluation metrics	Difficult comparison and validation	Researchers, Practitioners	Standardized benchmarks, Evaluation protocols

TRUST MECHANISMS AND USER ACCEPTANCE OF TRANSPARENT AI SYSTEMS

Psychological Factors Influencing Trust in AI Explanations

Psychological factors play a fundamental role in determining user acceptance and trust in explainable AI systems. Cognitive biases such as confirmation bias and availability heuristic influence how individuals interpret and respond to AI explanations, often leading users to accept explanations that align with their preexisting beliefs while rejecting contradictory information. Individual differences in risk tolerance, technical expertise, and domain knowledge significantly affect trust formation and calibration. The perceived competence, benevolence, and integrity of AI systems shape user willingness to rely on algorithmic recommendations, with explanations serving as crucial signals for establishing these trust dimensions.

User Studies on Explanation Effectiveness and Preference

Empirical research on user preferences and explanation effectiveness reveals significant variation in how different stakeholder groups respond to various explanation formats and techniques (Rong, Y. *et al*., 2023). Domain experts typically prefer detailed, technical explanations that align with their professional knowledge, while general users favor simplified, intuitive presentations that minimize cognitive load. Cultural factors and individual personality traits influence explanation preferences, with some users seeking comprehensive information while others prefer concise summaries. The effectiveness of explanations depends heavily on the match between explanation format and user characteristics, highlighting the need for personalized explanation systems.

Designing Human-Centered Explanation Interfaces

Human-centered design principles emphasize the importance of tailoring explanation interfaces to user needs, capabilities, and contexts. Effective explanation interfaces incorporate progressive disclosure mechanisms that allow users to access increasing levels of detail based on their interest and expertise. Interactive visualization techniques enable users to explore model behavior through direct manipulation and dynamic feedback, enhancing understanding and engagement. The design of explanation interfaces must balance information richness with cognitive accessibility, providing sufficient detail for informed decision-making while avoiding information overload that can impair user performance.

Professional Adoption Patterns Across Different Domains

Adoption patterns for explainable AI systems vary significantly across professional domains, reflecting differences in regulatory requirements,

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

risk tolerance, and existing decision-making processes. Healthcare professionals demonstrate cautious adoption patterns, emphasizing the need for explanations that align with clinical reasoning and established medical protocols. Financial sector professionals show greater acceptance of algorithmic explanations when they provide clear audit trails and comply with regulatory requirements. Legal professionals require explanations that can withstand judicial scrutiny and support due process requirements, leading to a preference for rule-based and precedent-aligned explanation formats.

Trust Calibration and Appropriate Reliance on AI Systems

Trust calibration represents the critical challenge of achieving appropriate levels of reliance on AI systems, avoiding both over-trust and under-trust that can lead to suboptimal outcomes (Schmid, A., & Wiesche, M. 2023). Effective trust calibration requires explanations that accurately communicate model uncertainty, limitations, and boundary conditions to help users make informed decisions about when to rely on algorithmic recommendations. The development of appropriate reliance patterns depends on user education, experience with the system, and feedback mechanisms that reinforce accurate mental models of AI capabilities and limitations.

Organizational and Institutional Trust-Building Strategies

Organizational adoption of explainable AI requires comprehensive trust-building strategies that address technical, procedural, and cultural factors. Institutional policies and governance frameworks provide the foundation for responsible AI deployment, establishing clear guidelines for explanation requirements and accountability mechanisms. Training programs and change management initiatives help organizations develop the capabilities and culture necessary for effective explainable AI adoption. Stakeholder engagement processes ensure that explanation systems meet the diverse needs of different user groups while maintaining organizational objectives and values.

Table 4: User Acceptance Factors and Trust Mechanisms Across Professional Domains (Rong, Y. *et al*., 2023; Schmid, A., & Wiesche, M. 2023)

Professional Domain	User Characteristics	Preferred Explanation Format	Trust Factors	Adoption Barriers	Trust-Building Strategies
Healthcare	High expertise, Risk-averse	Clinical reasoning alignment	Safety validation, Peer acceptance	Liability concerns, Workflow integration	Professional training, Gradual deployment
Finance	Regulation-focused, Performance-driven	Audit trails, Quantitative metrics	Compliance assurance, Performance consistency	Regulatory uncertainty, Cost concerns	Regulatory guidance, Industry standards
Legal	Precedent-oriented, Accountability-focused	Rule-based explanations	Due process alignment, Precedent consistency	Constitutional concerns, Bias detection	Legal framework development, Transparency requirements
General Users	Variable expertise, Convenience-seeking	Simple, intuitive presentations	Ease of use, Perceived benefit	Complexity, Trust deficit	User education, Progressive disclosure
Technical Experts	High technical knowledge, Detail-oriented	Comprehensive technical details	Methodological rigor, Reproducibility	Performance trade-offs, Implementation complexity	Technical documentation, Peer validation

CONCLUSION

The evolution of explainable artificial intelligence represents a critical advancement in addressing the transparency deficit inherent in contemporary machine learning systems. As AI applications proliferate across high-stakes domains including healthcare, finance, criminal justice, and autonomous systems, the demand for interpretable

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

and trustworthy algorithmic decision-making intensifies. The diverse taxonomies and methodologies for explainable AI, ranging from model-agnostic techniques to intrinsically interpretable architectures, provide multiple pathways for achieving transparency while maintaining predictive performance. However, significant technical challenges persist, including the accuracy-interpretability trade-off, computational scalability constraints, explanation fidelity concerns, and vulnerability to adversarial manipulation. Human factors emerge as equally important considerations, with cognitive limitations and psychological biases influencing explanation effectiveness and user acceptance patterns. Trust mechanisms and calibration strategies prove essential for fostering appropriate reliance on transparent AI systems across different professional domains. The successful deployment of explainable AI requires interdisciplinary collaboration that addresses both technical capabilities and human-centered design principles. Future developments must prioritize standardized evaluation frameworks, domain-specific interpretability solutions, and robust governance mechanisms to realize the full potential of trustworthy artificial intelligence in society. The convergence of technical innovation with ethical imperatives positions explainable AI as a cornerstone for responsible algorithmic deployment in an increasingly automated world.

REFERENCES

1. Ortigossa, E. S., Gonçalves, T., & Nonato, L. G. "Explainable artificial intelligence (xai)—from theory to methods and applications." *IEEE Access* 12 (2024): 80799-80846.
2. Rendón, L. G. "An introduction to the principle of transparency in automated decision-making systems." *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, (2022).
3. Ma, Y., Wang, Z., Yang, H., & Yang, L. "Artificial intelligence applications in the development of autonomous vehicles: A survey." *IEEE/CAA Journal of Automatica Sinica* 7.2 (2020): 315-329.
4. Sushma, M., Goud, B. Y., Nikhil, M., & Reddy, G. S. K. "AI Medical Diagnosis Application." *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, (2021)
5. Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N., & Narvekar, M. "A study of LIME and SHAP model explainers for autonomous disease predictions." *2022 ieee bombay section signature conference (ibssc)*. IEEE, (2022)
6. Islam, M. R., Ahmed, M. U., & Begum, S. "Local and global interpretability using mutual information in explainable artificial intelligence." *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, (2021)
7. Griffiths, T. L. "Understanding human intelligence through human limitations." *Trends in Cognitive Sciences* 24.11 (2020): 873-883.
8. Jeanneret, G., Simon, L., & Jurie, F. "Adversarial counterfactual visual explanations." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023)
9. Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., ... & Kasneci, E. "Towards human-centered explainable ai: A survey of user studies for model explanations." *IEEE transactions on pattern analysis and machine intelligence* 46.4 (2023): 2104-2122.
10. Schmid, A., & Wiesche, M, "The importance of an ethical framework for trust calibration in AI." *IEEE Intelligent Systems* 38.6 (2023): 27-34.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Taneja, A. "Explainable AI: Building Trust and Transparency in Machine Learning Models" *Sarcouncil Journal of Engineering and Computer Sciences* 4.8 (2025): pp 482-489.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher