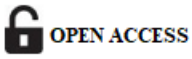
International Journal of Computer Scientist (IJCSCT)
Volume 2, Issue 1, January-June (2025), pp. 1-7, Article ID: IJCSCT_02_01_001
Available online at https://iaeme.com/Home/issue/IJCSCT?Volume=2&Issue=1
Journal ID: 8381-9009



EXPLAINABLE ARTIFICIAL INTELLIGENCE MODELS FOR TRANSPARENT AND TRUSTWORTHY DECISION SUPPORT SYSTEMS

Alex Fernando
AI Governance Analyst, Spain.

ABSTRACT

Decision support systems increasingly rely on artificial intelligence to enhance accuracy, efficiency, and automation across domains such as healthcare, finance, engineering, and governance. However, the opacity of many advanced artificial intelligence models has raised concerns regarding transparency, accountability, and trust. Explainable artificial intelligence models address these concerns by providing interpretable insights into model behavior and decision logic. This paper examines the role of explainable artificial intelligence in developing transparent and trustworthy decision support systems. It presents conceptual foundations, modeling approaches, system architectures, and evaluation metrics that support explainability.

Keywords: Explainable Artificial Intelligence, Decision Support Systems, Model Transparency, Trustworthy AI, Interpretability, Responsible AI

Cite this Article: Alex Fernando. (2025). Explainable Artificial Intelligence Models for Transparent and Trustworthy Decision Support Systems. *International Journal of Computer Scientist (IJCSCT),* 2(1), 1-7.

https://iaeme.com/MasterAdmin/Journal_uploads/IJCSCT/VOLUME_2_ISSUE_1/IJCSCT_02_01_001.Pdf

1. Introduction

Decision support systems play a critical role in assisting human decision-makers by analyzing data and recommending actions. With the integration of artificial intelligence, these systems can process complex data patterns and generate highly accurate predictions. However, many artificial intelligence models operate as black boxes, limiting user understanding of how decisions are produced.

The lack of transparency in artificial intelligence-driven decision support systems poses risks related to trust, accountability, and governance. Users may hesitate to rely on recommendations they cannot interpret, particularly in high-stakes applications. Explainable artificial intelligence has therefore emerged as a key requirement for responsible deployment.

This paper explores explainable artificial intelligence models as enablers of transparent and trustworthy decision support systems. It focuses on conceptual foundations, system architectures, evaluation mechanisms, and practical implications for organizations.

2. Conceptual Foundations of Explainable Artificial Intelligence

Explainable artificial intelligence refers to methods and models that enable humans to understand, interpret, and trust artificial intelligence decisions. These methods aim to make model behavior transparent without significantly compromising predictive performance. Explainability is closely related to concepts such as interpretability, accountability, and fairness.

In decision support systems, explainability serves both technical and organizational purposes. Technically, it supports model validation and debugging. Organizationally, it enhances user confidence and supports informed decision-making. Explainable models bridge the gap between complex algorithms and human reasoning.

The foundation of explainable artificial intelligence emphasizes human-centered design. Models are developed not only for accuracy but also for clarity, justification, and communicability of decisions.

3. Literature Review

3.1 Foundational Work on Explainability

Ribeiro, Singh, and Guestrin (2016) introduced LIME (Local Interpretable Model-agnostic Explanations), a method designed to explain individual predictions of any machine learning classifier by approximating it locally with an interpretable surrogate model. LIME's

approach is critical in providing transparency for black-box models by making them more understandable for users in practical decision-making environments. The method has been widely adopted due to its simplicity and flexibility in application across different machine learning models (Ribeiro et al., 2016).

Lundberg and Lee (2017) developed SHAP (SHapley Additive exPlanations), a unified framework based on game theory that attributes feature importance for any machine learning model. SHAP values provide a consistent approach to explain individual predictions by assigning an importance score to each feature, ensuring that the explanation is both accurate and fair. SHAP's foundation in cooperative game theory ensures that the feature contributions are calculated in a manner that reflects their impact on model output, making it a powerful tool for interpretability (Lundberg & Lee, 2017).

3.2 XAI for Decision Support Systems

Caruana et al. (2015) compared rule-based models with black-box models in the context of clinical decision support systems (CDSS). They demonstrated that rule-based models not only maintained high predictive performance but also provided clear and actionable explanations, which are essential in medical applications. The transparency of these models made them highly valuable in practice, particularly in healthcare settings where understanding the reasoning behind decisions is critical for clinicians. Their work highlighted the need for combining interpretability with accuracy in high-stakes decision-making environments (Caruana et al., 2015).

Guidotti et al. (2018) proposed a taxonomy of explanation methods and reviewed their applications in various domains, including fraud detection and risk assessment systems. They emphasized that different application areas require different explanation methods, ranging from local explanations for specific predictions to global insights that help understand model behavior over time. Their taxonomy helps clarify the various approaches to explaining AI models and demonstrates how they can be tailored to enhance the transparency and trustworthiness of decision support systems in diverse sectors (Guidotti et al., 2018).

4. Explainable AI Models and Techniques

Explainable artificial intelligence models include inherently interpretable models and post-hoc explanation techniques. Interpretable models such as decision trees and rule-based systems provide transparent decision logic by design. These models are well-suited for applications where traceability is essential.

Post-hoc techniques generate explanations for complex models after decisions are made. These methods approximate local or global model behavior and highlight influential features. They allow advanced models to be used while still providing insight into their reasoning.

The choice of explainable model depends on system requirements, data complexity, and regulatory expectations. Decision support systems often combine multiple explainability techniques to enhance clarity and robustness.

Table 1: Categories of Explainable AI Models

Model Category	Explanation Approach
Interpretable Models	Transparent decision logic
Post-hoc Explainers	Feature influence analysis
Hybrid Models	Combined transparency and accuracy

5. Architecture of Explainable Decision Support Systems

Explainable decision support systems integrate data processing, artificial intelligence models, and explanation layers into a unified architecture. The explanation layer translates model outputs into human-understandable insights. This layer is critical for trust and accountability.

System architecture typically separates model computation from explanation generation. This separation allows flexibility in updating explanation techniques without modifying core models. It also supports scalability across applications.

By embedding explainability into system architecture, organizations ensure consistent transparency across decision workflows. This integration enhances system governance and user engagement.

6. Visualization and Interaction for Explainability

Visualization plays a central role in communicating explanations to users. Graphs, feature importance charts, and decision paths make abstract model logic accessible. Effective visualization reduces cognitive load and improves comprehension.

Interactive explanation interfaces allow users to explore alternative scenarios and model sensitivities. This interactivity supports deeper understanding and informed decision-making. Visualization thus transforms explanations into actionable insights.

Explainable decision support systems rely on well-designed visualization to bridge technical complexity and human interpretation. Poor visualization can undermine explainability despite robust underlying models.

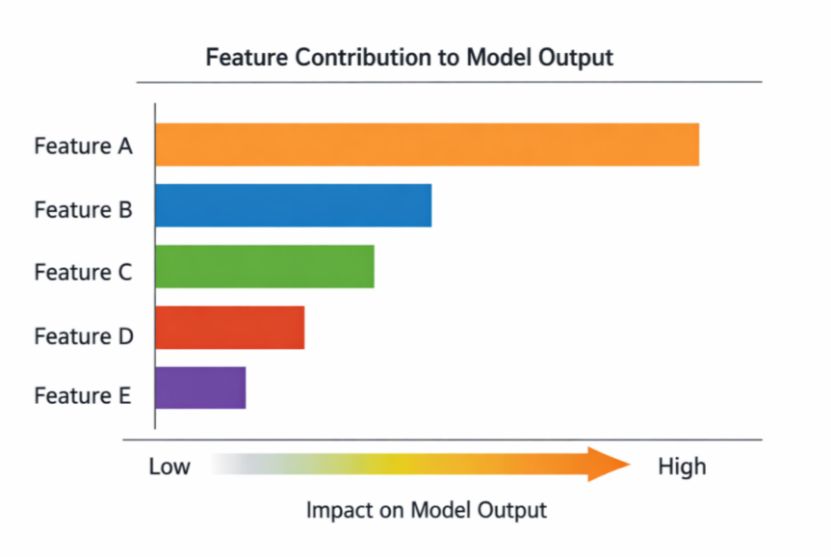


Figure 1: Feature Contribution Visualization for Explainability

7. Trust, Accountability, and Governance

Trust is a foundational requirement for decision support systems. Explainable artificial intelligence enhances trust by making decisions understandable and justifiable. Users are more likely to accept recommendations when reasoning is transparent.

Accountability is strengthened through explainability by enabling traceability of decisions. Organizations can identify responsible components and actors within decision processes. This supports internal governance and external compliance.

Governance frameworks increasingly require transparency in artificial intelligence systems. Explainable models align with these requirements by embedding accountability and oversight into system design.

8. Performance Metrics for Explainable Decision Support

Evaluating explainable decision support systems requires metrics beyond predictive accuracy. Explanation quality, consistency, and user comprehension are critical evaluation dimensions. These metrics assess whether explanations truly support understanding.

Trade-offs often exist between explainability and performance. Systems must balance interpretability with analytical power. Performance evaluation therefore considers both technical and human-centered outcomes.

Comprehensive evaluation supports continuous improvement of explainable systems. Metrics guide refinement of models and explanation techniques.

Table 2: Evaluation Metrics for Explainable AI

Metric	Purpose
Prediction Accuracy	Decision quality
Explanation Fidelity	Alignment with model
User Comprehension	Interpretability assessment

9. Conclusion

Explainable artificial intelligence models are essential for developing transparent and trustworthy decision support systems. They address critical challenges related to opacity, accountability, and trust in artificial intelligence-driven decisions. By integrating explainable models, visualization, and governance mechanisms, organizations can enhance both decision quality and user confidence. The proposed frameworks illustrate how explainability can be embedded across system architectures. Despite technical and organizational challenges, the benefits of explainable artificial intelligence are substantial. These models support responsible innovation and regulatory alignment. Overall, explainable artificial intelligence represents a foundational component of trustworthy decision support systems.

References

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

[2] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.

[3] Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.

[4] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD Conference.

[5] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems.

[6] Molnar, C. (2019). Interpretable machine learning. Leanpub.

[7] Binns, R. (2018). Fairness in machine learning. Communications of the ACM, 61(11), 86–94.

[8] Burrell, J. (2016). How the machine "thinks". Big Data & Society, 3(1), 1–12.

[9] Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62.

[10] Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework. Minds and Machines, 28(4), 689–707.

[11] Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705.

[12] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms. Big Data & Society, 3(2), 1–21.

[13] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing. Proceedings of the AAAI Conference on AI Ethics.

[14] Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making with machine learning. Academy of Management Review, 44(1), 66–87.

Citation: Alex Fernando. (2025). Explainable Artificial Intelligence Models for Transparent and Trustworthy Decision Support Systems. International Journal of Computer Scientist (IJCSCT) 2(1), 1-7.

Abstract Link: https://iaeme.com/Home/article_id/IJCSCT_02_01_001

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJCSCT/VOLUME_2_ISSUE_1/IJCSCT_02_01_001.pdf

✉ **editor@iaeme.com**