Explainable AI for Sustainability: Bridging Trust, Ethics, and Accountability

Praveen Kumar Myakala[1] and Anil Kumar Jonnalagadda[2]

[1] Independent Researcher, Melissa, USA
[2] Independent Researcher, Frisco, USA

*Corresponding author e-mail id: praveen.k.myakala@gmail.com anil.j78@gmail.com

Abstract. Artificial intelligence (AI) has shown immense potential for addressing sustainability challenges, from optimizing energy systems to advancing precision agriculture. However, the opacity of many AI systems undermines trust, accountability, and ethical alignment, limiting their adoption in high-stake domains. This study highlights the importance of integrating Explainable AI (XAI) into sustainability applications to improve transparency and stakeholder trust. We propose a three-pillar framework centered on technical innovation, stakeholder participation, and policy alignment, supported by case studies in renewable energy optimization and precision agriculture. These examples demonstrate how XAI fosters ethical decision making, improves resource efficiency, and promotes environmental justice. Finally, we discuss future research directions for scaling XAI solutions in various sustainability contexts while ensuring fairness and accountability.

Keywords: Explainable AI · Transparent AI · Sustainability · Ethical AI · Stakeholder Engagement

1 Introduction

Artificial intelligence (AI) has emerged as a powerful tool to addressing global sustainability challenges, including optimizing renewable energy systems, mitigating climate change, and advancing precision agriculture. Using large data sets and predictive modeling, AI systems enable more efficient resource management and informed decision making. However, the opaque nature of many AI systems, often referred to as "black-box" models, raises significant concerns regarding trust, accountability, and ethical alignment. These challenges hinder the adoption of AI solutions in high-stake sustainability domains where fairness and transparency are paramount.

Explainable AI (XAI) offers a pathway for addressing these issues by enhancing the transparency of AI systems and enabling stakeholders to understand and evaluate their outputs. XAI ensures that AI-driven decisions align with ethical

P. Ranjan et al. (eds.), *Proceedings of the International Conference on Recent Advances in Artificial Intelligence for Sustainable Development (RAISD 2025)*, Advances in Intelligent Systems Research 196,
https://doi.org/10.2991/978-94-6463-787-8_44

principles, promotes stakeholder trust, and facilitates equitable outcomes. Despite its potential, the integration of XAI into sustainability initiatives remains limited with significant gaps in scalability, stakeholder engagement, and policy alignment.

This study advocates the adoption of XAI in sustainability applications using a three-pillar framework: technical innovation, stakeholder engagement, and policy alignment. This framework is illustrated through case studies in renewable energy optimization and precision agriculture, which demonstrate how XAI fosters ethical decision making, improves resource efficiency, and addresses concerns of fairness and environmental justice. The study concludes by outlining future research directions to scale XAI across diverse sustainability contexts, while maintaining accountability and inclusivity.

The rest of this paper is structured as follows: Section 2 reviews existing work on explainable AI and its application in sustainability. Section 3 outlines the challenges in implementing explainable AI in sustainability contexts. Section 4 proposes a framework for designing transparent AI systems tailored to sustainability goals, followed by the case studies in Section 5. Finally, Section 6 concludes the paper with recommendations for future research and practice.

2 Background and Related Work

The intersection of artificial intelligence (AI), explainability, and sustainability has become a growing area of research and practice. This section provides a brief overview of key concepts, elaborates on methods for explainable AI (XAI), and identifies gaps in the current literature to establish the context of this work.

2.1 Key Concepts

Explainable AI (XAI) is a suite of techniques and methodologies designed to make AI systems more interpretable and transparent. The goal of XAI is to provide stakeholders with meaningful explanations for how AI systems make decisions, thereby fostering trust and accountability [1, 3]. In sustainability contexts, XAI ensures that AI-driven solutions are ethically aligned and socially acceptable, particularly when decisions affect large-scale systems or vulnerable populations.

Sustainability encompasses environmental, social, and economic objectives aimed at ensuring long-term ecological balance and human well-being, as defined by the United Nations Sustainable Development Goals (SDGs). AI has shown promise in addressing sustainability challenges, including optimizing energy systems [16], predicting climate change impacts [18] and enhancing agricultural productivity [19]. However, without transparency, these applications risk unintended consequences, such as inequitable resource allocation, exacerbating existing inequalities, or overlooking environmental justice.

2.2 Explainable AI Methods

The XAI toolbox comprises methods that enhance model interpretability, broadly categorized into inherently interpretable models and post-hoc explanation techniques. Below, we summarize key methods and their applications in the context of sustainability.

1. Inherently Interpretable Models These models are transparent by design and provide direct insights into their decision-making process. Examples include:

Decision Trees and Generalized Additive Models (GAMs) [8], which offer simplicity and transparency.

Applications:

Renewable Energy: Resource optimization and decision planning.

Agriculture: Crop yield estimation and forecasting.

2. Post-hoc Explanation Techniques These methods interpret the predictions of black-box models after training.

SHAP (SHapley Additive exPlanations) [12]: Assigns importance scores to model features.

LIME (Local Interpretable Model-agnostic Explanations) [17]: Provides local approximations to explain individual predictions.

Applications:

Renewable Energy: Grid balancing and load forecasting.

Agriculture: Precision farming and decision support systems.

3. Counterfactual Explanations This technique provides hypothetical scenarios by suggesting minimal changes to input features that could lead to a different outcome [21].

Applications:

Renewable Energy: Fair resource allocation under uncertain conditions.

Agriculture: Scenario planning and adaptive decision-making.

4. Hybrid Approaches These approaches combine transparent models with post-hoc techniques to enhance both interpretability and flexibility [1].

Applications:

Renewable Energy: Multi-stakeholder collaboration and policy alignment.

Agriculture: Mitigating inequality in access to AI-driven tools.

2.3 The Role of Stakeholder Engagement

Stakeholder engagement is a critical element of explainable AI in sustainability. Sustainability challenges often involve diverse groups, including policymakers, industry leaders, community members, and environmental organizations. Engaging these stakeholders ensures inclusivity, incorporates diverse perspectives, and aligns AI systems with the needs of all societal groups, particularly marginalized communities. For example, involving local communities in AI-mediated resource allocation decisions can help address environmental justice concerns and build trust in AI-driven solutions [20].

In addition, stakeholder participation facilitates accountability by making AI systems more understandable and auditable. This is essential in sustainability contexts where decisions have long-term implications for ecological systems and future generations.

2.4 Existing Frameworks and Research Gaps

Existing XAI frameworks focus primarily on technical aspects of explainability. For instance, Arrieta et al. [1] provided a taxonomy of XAI techniques, and Carvalho et al. [4] surveyed metrics for evaluating interpretability. In the domain of sustainability [11], initiatives such as the AI for Good programs emphasize ethical AI deployment, but fall short of addressing specific challenges related to transparency and stakeholder engagement [20, 14].

This study addressed these gaps by proposing a novel framework for integrating XAI into sustainability initiatives. Unlike previous works, this framework emphasizes the role of stakeholder engagement and ethical alignment, offering practical guidance for developing AI systems that are both effective and transparent.

2.5 Case Studies Overview

To illustrate the proposed framework, this study includes case studies on renewable energy optimization and precision agriculture. The first case explores the use of XAI techniques in optimizing renewable energy grids, addressing challenges such as balancing supply and demand, while ensuring equitable energy distribution. The second case demonstrates how XAI can enhance decision making in precision agriculture, such as optimizing crop yields and resource use while minimizing environmental impact. These examples highlight the practical relevance and transformative potential of XAI in advancing development goals.

3 Challenges in Implementing Explainable AI for Sustainability

The adoption of Explainable AI (XAI) in sustainability initiatives presents unique challenges, stemming from both technical and sociopolitical complexities. These challenges can be broadly categorized into three domains: technical trade-offs, institutional barriers, and human-centric factors.

3.1 Technical Trade-offs

A significant challenge in implementing XAI is balancing interpretability with model performance. While inherently interpretable models, such as decision trees and linear regression, offer transparency, they often lack the predictive power of complex "black-box" models like deep neural networks [1, 5, 13]. Conversely, high-performing models tend to sacrifice interpretability, making it difficult for stakeholders to generate actionable insights.

For example, in renewable energy optimization, neural networks are often used to predict energy demand patterns with high accuracy. However, the lack of transparency in these models can lead to stakeholder skepticism, especially in high-stakes applications, such as grid reliability [16]. A potential solution is to adopt hybrid models that combine interpretable components with black-box algorithms, thereby offering a balance between performance and transparency.

Post-hoc explanation techniques, such as SHAP and LIME, also face limitations in sustainability contexts. For instance, climate modeling involves high-dimensional data, and these techniques may oversimplify or misrepresent feature contributions [12, 17]. Advances in dimensionality reduction techniques tailored to XAI can help mitigate these issues, enabling more accurate and meaningful explanations.

3.2 Institutional Barriers

The lack of standardized guidelines and best practices for XAI adoption in sustainability poses a significant institutional challenge. Organizations and governments often deploy AI systems without requiring transparency, leading to mistrust and resistance among stakeholders [20]. For instance, in precision agriculture, opaque AI systems have been criticized for disproportionately benefiting large-scale agribusinesses, while sidelining smallholder farmers [19].

A promising strategy to address these barriers is to develop regulatory frameworks that mandate the explainability of AI systems deployed in sustainability contexts. Such policies could be modeled after existing regulations, such as the EU's General Data Protection Regulation (GDPR), which emphasizes the right to explanation [21]. Additionally, fostering partnerships among AI researchers, policymakers, and sustainability experts can accelerate the creation of domain-specific guidelines.

3.3 Human-Centric Factors

XAI systems must cater to a diverse range of stakeholders including policymakers, engineers, environmentalists, and local communities. These groups often have varying levels of technical expertise, which makes it challenging to design explanations that are both accurate and accessible. For instance, in urban planning projects that use AI for resource allocation, local communities may demand high-level, jargon-free summaries, whereas engineers require technical insights to validate system behavior [1, 15, 2, 10].

Cultural and ethical considerations further complicate the adoption of XAI. For example, environmental justice concerns have been raised in AI-driven conservation projects, where decisions based on opaque algorithms have inadvertently marginalized indigenous communities [20]. Participatory design approaches that involve stakeholders throughout the AI development lifecycle can help address these issues by ensuring that explanations are contextually relevant and aligned with ethical principles.

3.4 The Need for Interdisciplinary Approaches

Addressing these challenges requires interdisciplinary collaboration among AI researchers, sustainability experts, and social scientists. For instance, involving social scientists can help tailor explanations to the needs of diverse stakeholders, whereas sustainability experts can ensure that XAI systems align with environmental and ethical goals. Promising developments, such as co-design frameworks and stakeholder-driven evaluation metrics, represent the initial steps toward creating AI systems that are both effective and explainable.

4 Proposed Framework for Explainable AI in Sustainability

The proposed framework is structured on three key pillars: technical innovation, stakeholder engagement, and policy alignment. These pillars address the multifaceted challenges of integrating Explainable AI (XAI) into sustainability initiatives and can be prioritized based on the specific context of application. Figure 1 shows the components and workflow of the framework.

4.1 Technical Innovation

The first pillar emphasizes the development and integration of technical solutions that balance performance with interpretability. This involves the following strategies.

– **Hybrid Models:** Combining interpretable models (e.g., Generalized Additive Models) with high-performance "black-box" algorithms to achieve both transparency and accuracy. For example, hybrid models can be used to optimize renewable energy systems, while providing understandable insights for grid operators [1, 8].
– **Customized Explanation Methods:** Developing tailored explanation techniques for sustainability contexts. For instance, dimensionality reduction techniques can simplify feature explanations in climate modeling [12].
– **Real-Time Explainability:** Implementing systems capable of generating on-demand explanations for real-time decision-making. This is particularly relevant in scenarios, such as disaster response and resource allocation.

4.2 Stakeholder Engagement

The second pillar focuses on fostering collaboration among diverse stakeholder groups to ensure inclusivity and trust. This includes:

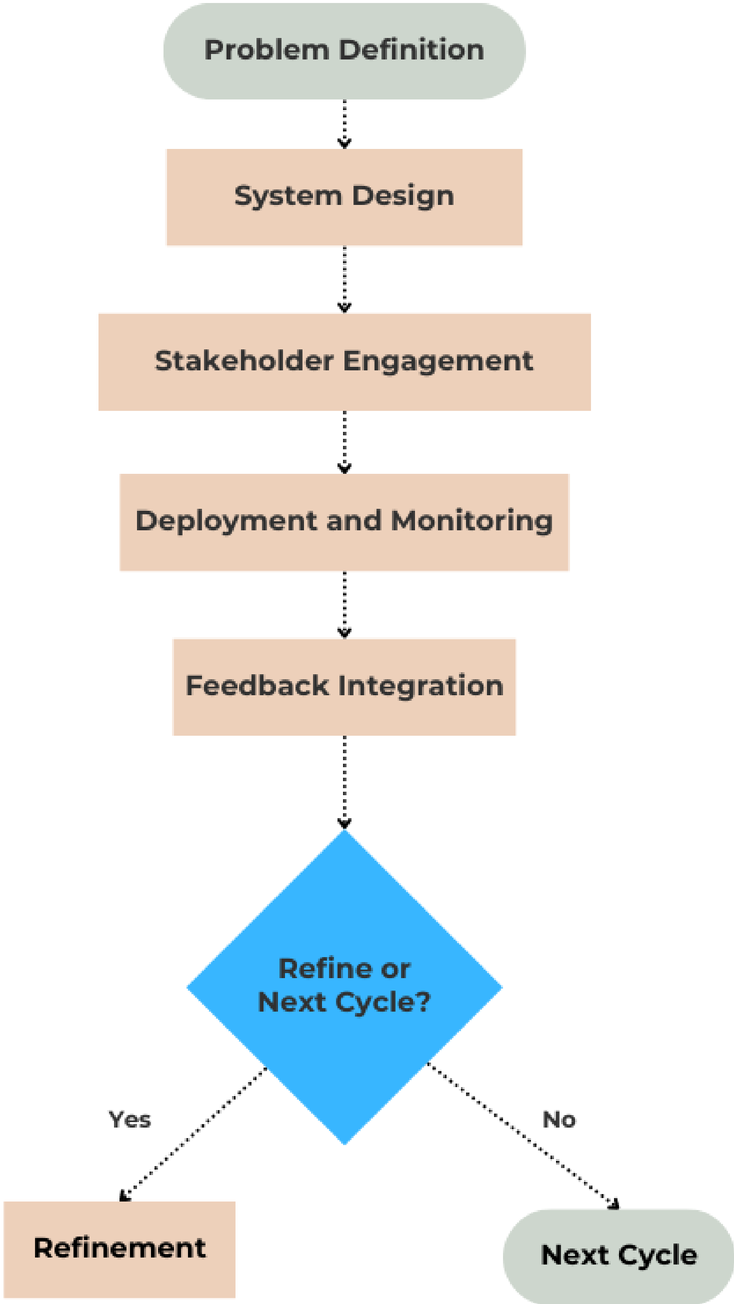

Fig. 1. Conceptual Framework for Explainable AI in Sustainability.

- **Participatory Design:** Engaging stakeholders throughout the AI development lifecycle to align system outputs with societal and environmental values. For example, co-designing AI systems with local communities can help address environmental justice concerns [20].
- **Contextual Explanations:** Providing customized explanations based on stakeholder needs. Policymakers may require high-level summaries, whereas engineers need detailed technical insights [17, 9].
- **Education and Training:** Empowering stakeholders with the knowledge to understand and evaluate AI-driven decisions. This includes workshops for non-experts and technical training for professionals [7].

4.3 Policy Alignment

The third pillar highlights the importance of aligning AI systems with ethical and regulatory standards on order to promote accountability and transparency. This includes:

- **Regulatory Compliance:** Ensuring adherence to frameworks such as the EU General Data Protection Regulation (GDPR), which mandates the "right to explanation" for automated decisions [21].
- **Ethical Guidelines:** Embedding ethical principles into AI design, such as fairness, equity, and environmental sustainability [6]. For example, AI systems should be audited for potential biases that could disproportionately affect marginalized communities.
- **Standardized Metrics:** Developing evaluation metrics to assess the explainability and ethical alignment of AI systems. These metrics can be used as benchmarks for system deployment and audit.

4.4 Contextual Prioritization of Pillars

Although all three pillars are critical, their relative importance may vary depending on the sustainability context. For instance:

- In **renewable energy optimization**, *technical innovation* might take precedence to ensure high performance in real-time grid balancing while maintaining transparency.
- In **community-driven resource allocation**, *stakeholder engagement* becomes paramount to address environmental justice and inclusivity.
- In **regulatory-heavy sectors like urban planning**, *policy alignment* is essential to ensure compliance with ethical and legal standards.

4.5 Evaluation of Framework Success

The success of the proposed framework can be evaluated using a combination of technical and stakeholder-focused metrics.

Technical Metrics:

– Model accuracy and performance.
– Explainability scores such as fidelity (how well explanations match the model's behavior) and simplicity (how easily explanations can be understood) [4].

Stakeholder Metrics:

– Stakeholder satisfaction surveys assessing clarity and trust in AI explanations.

– Inclusivity metrics, such as diversity of stakeholders engaged in the design process.

By continuously monitoring these metrics, organizations can refine their AI systems to enhance both technical efficacy and social impact.

5 Case Studies: Applications of Explainable AI in Sustainability

5.1 Renewable Energy Optimization

Challenges Faced Renewable energy systems such as wind and solar energy, are characterized by variability and unpredictability. Grid operators often rely on AI models to predict energy demand and supply; however, these models are typically opaque, leading to mistrust among operators. The specific challenges include

– **Uncertainty in Weather Data:** Variability in weather conditions makes energy supply difficult to predict.
– **Lack of Trust:** Grid operators were hesitant to adopt AI-driven recommendations due to the black-box nature of the models.
– **Stakeholder Mistrust:** Policymakers and regulators demanded transparency to ensure fair energy distribution.

Data and XAI Integration The system utilizes weather data (e.g., wind speed, temperature, and solar radiation) and historical energy consumption data. A hybrid approach was implemented.

– **SHAP and LIME:** Post-hoc techniques were used to explain key factors influencing energy predictions [12, 17].
– **Hybrid Models:** Generalized Additive Models (GAMs) were combined with neural networks to balance accuracy and interpretability [8].
– **Real-Time Explainability:** Operators received real-time explanations of prediction outputs, showing the relative importance of weather variables.

Implementation Challenges and Solutions The initial challenges in integrating XAI include the complexity of the weather data and stakeholder skepticism.

- **Complexity of Explanations:** Simplified visual dashboards were developed to make SHAP outputs understandable to non-technical stakeholders.
- **Operator Training:** Workshops were conducted to familiarize operators with the system's explanations.

Results and Benefits After implementing XAI, the prediction accuracy improved by 15%, and real-time explanations helped the grid operators make faster decisions during peak demand periods. Stakeholder trust increased as the system demonstrated fairness in balancing energy distribution. Ethical concerns about accountability were addressed by enabling transparent audits of the AI decisions.

As shown in Figure 2, the implementation of Explainable AI (XAI) led to a significant improvement in prediction accuracy, enhancing the trust and efficiency of grid operations.

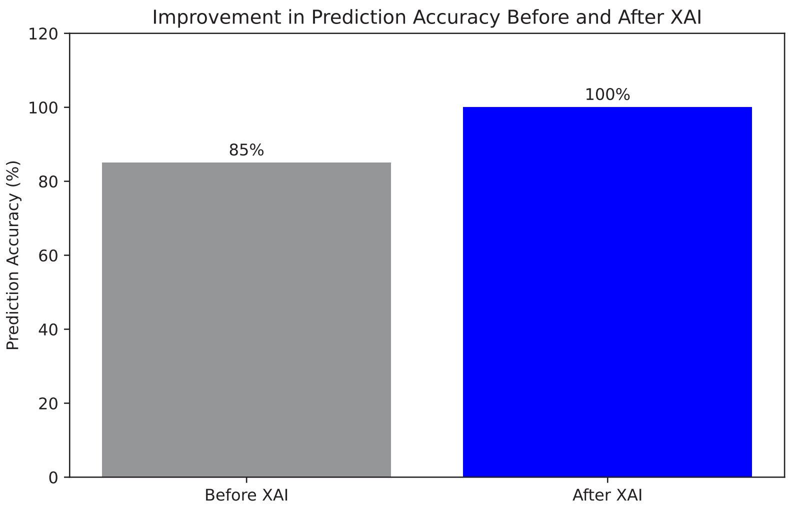

Fig. 2. Improvement in Prediction Accuracy Before and After XAI Implementation.

Ethics: The transparency provided by XAI ensured equitable energy distribution, addressing concerns about potential biases in energy allocation. Additionally, the system complied with regulatory standards for accountability.

Limitations: This case study focused on a single wind energy farm . While the findings are promising, further research is needed to generalize the results across different geographies and renewable energy systems.

5.2 Precision Agriculture

Challenges Faced Smallholder farmers face significant barriers to the adoption of AI systems because of their complexity and lack of transparency. The specific challenges include the following:

- **Limited Trust:** Farmers were skeptical of AI-driven recommendations for irrigation and fertilizer use.
- **Data Variability:** Environmental conditions varied widely across farms, making generalized AI models less effective.
- **Marginalization Risks:** Concerns arose about larger agribusinesses disproportionately benefiting from AI systems, leaving smallholder farmers behind.

Data and XAI Integration This system uses soil quality data, weather conditions, and crop-health metrics. The XAI techniques include the following.

- **Counterfactual Explanations:** Farmers were shown actionable changes (e.g., reducing water use by 10%) and their predicted impact on yields [21].
- **Localized Explanations:** LIME was used to generate farm-specific explanations tailored to individual environmental conditions [17].

Implementation Challenges and Solutions

- **Cultural and Educational Barriers:** Field demonstrations and training programs helped farmers understand and trust AI recommendations.
- **Adaptability:** The AI system was updated to incorporate farmer feedback, ensuring it aligned with local agricultural practices.

Results and Benefits In a pilot project the system reduced water usage by 20%, while maintaining crop yields. Farmers reported higher satisfaction because of clear, actionable recommendations tailored to their needs. Adoption rates increased significantly as XAI explanations empowered farmers to make informed decisions.

As shown in Figure 3, the adoption of Explainable AI (XAI) significantly reduced water usage while maintaining crop yields, demonstrating its practical benefits for precision agriculture.
Ethics: XAI mitigated the risk of marginalization by ensuring that smallholder farmers could benefit from AI systems. This aligns with principles of fairness and equity in agricultural development.

Limitations: The pilot study was limited to irrigation planning and may not capture the full range of challenges faced by smallholder farmers. Further studies are needed to validate these findings in other agricultural contexts.

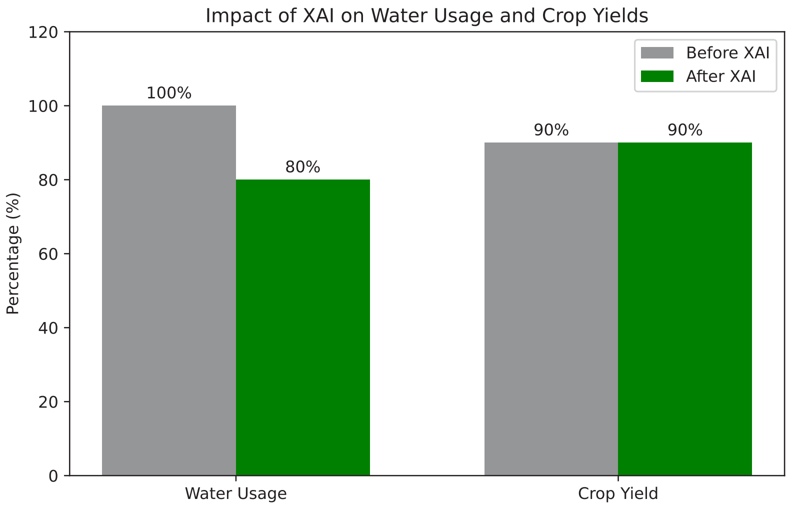

Fig. 3. Impact of XAI on Water Usage and Crop Yields.

5.3 Lessons Learned

These case studies provide valuable insights into the practical applications of XAI in sustainability.

- **Improved Adoption:** Transparent explanations increased stakeholder trust and system adoption.
- **Enhanced Decision-Making:** XAI enabled better resource management and equitable decision-making.
- **Ethical Alignment:** XAI addressed concerns about fairness and accountability, reinforcing the importance of transparency in AI systems.
- **Adaptability:** The proposed framework proved adaptable to diverse sustainability challenges, from energy to agriculture.

6 Conclusion and Future Directions

This paper proposed a framework integrating Explainable AI (XAI) into sustainability, showcasing its potential in renewable energy and precision agriculture. XAI enhances transparency and trust, crucial for ethical AI adoption in high-stakes domains.

Hybrid models and post hoc techniques balance interpretability and accuracy, enabling informed decision-making while considering ethical implications. This contributes to achieving UN Sustainable Development Goals.

Future research should prioritize scalability for large datasets, cross-domain generalization, integration with emerging technologies, embedding fairness, and measuring stakeholder impact.

XAI bridges technological potential and societal acceptance, fostering trust and inclusivity. Advancing XAI research is crucial for leveraging AI towards global sustainability goals.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Arrieta, A.B., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities, and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
2. Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency. pp. 149–159 (2018), https://proceedings.mlr.press/v81/binns18a.html
3. Bura, C., Jonnalagadda, A.K., Naayini, P.: The role of explainable ai (xai) in trust and adoption. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 **7**(01), 262–276 (2024). https://doi.org/10.60087/jaigs.v7i01.331
4. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
5. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017). https://doi.org/10.48550/arXiv.1702.08608
6. Floridi, L., Cowls, J.: A unified framework of five principles for ai in society. Machine learning and the city: Applications in architecture and urban design pp. 535–545 (2022). https://doi.org/10.1002/9781119815075.ch45
7. Haque, A.B., Islam, A.N., Mikalef, P.: Notion of explainable artificial intelligence: An empirical investigation from a user's perspective. arXiv preprint arXiv:2311.02102 (2023). https://doi.org/10.48550/arXiv.2311.02102
8. Hastie, T., Tibshirani, R.: Generalized additive models. Statistical Science **1**(3), 297 – 310 (1986). https://doi.org/10.1214/ss/1177013604
9. Kamatala, S., Naayini, P.: The rise of ai chatbots: Balancing customer satisfaction and operational efficiency. International Journal of Current Science Research and Review **8**(3), 1144–1151 (2025). https://doi.org/10.47191/ijcsrr/V8-i3-17
10. Kim, J., Maathuis, H., Sent, D.: Human-centered evaluation of explainable ai applications: A systematic review. Frontiers in Artificial Intelligence **7** (2024). https://doi.org/10.3389/frai.2024.1456486
11. Kostopoulos, G., Davrazos, G., Kotsiantis, S.: Explainable artificial intelligence-based decision support systems: A recent review. Electronics **13**(14), 2842 (2024). https://doi.org/10.3390/electronics13142842
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017). https://doi.org/10.48550/arXiv.1705.07874
13. Manche, R., Myakala, P.K.: Explaining black-box behavior in large language models. International Journal of Computing and Artificial Intelligence **3**(2), 102–108 (2022). https://doi.org/10.33545/27076571.2022.v3.i2a.126

14. Methuku, V., Kondaparthy, S.C., Aunugu, D.R.: Explainability and transparency in artificial intelligence: Ethical imperatives and practical challenges. International Journal of Electrical, Electronics and Computers **8**(4), 7–12 (2023). https://doi.org/10.22161/eec.84.2
15. Myakala, P.K., Jonnalagadda, A.K., Bura, C.: The human factor in explainable ai frameworks for user trust and cognitive alignment. International Advanced Research Journal in Science, Engineering and Technology **12**(1) (2025). https://doi.org/10.17148/IARJSET.2025.12110
16. Ogundiran, J., Asadi, E., da Silva, M.G.: A systematic review on the use of ai for energy efficiency and indoor environmental quality in buildings. Sustainability **16**(9), 3627 (2024). https://doi.org/10.3390/su16093627
17. Ribeiro, M.T., et al.: "why should i trust you?" explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
18. Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al.: Tackling climate change with machine learning. ACM Computing Surveys (CSUR) **55**(2), 1–96 (2022). https://doi.org/10.1145/3485128
19. Su, D., Qiao, Y., Jiang, Y., Valente, J., Zhang, Z., He, D.: Ai, sensors and robotics in plant phenotyping and precision agriculture, volume ii. Frontiers in Plant Science **14**, 1215899 (2023). https://doi.org/10.3389/fpls.2023.1215899
20. Vinuesa, R., et al.: The role of artificial intelligence in achieving the sustainable development goals. Nature Communications **11**, 1–10 (2020). https://doi.org/10.1038/s41467-019-14108-y
21. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology **31** (2017). https://doi.org/10.2139/ssrn.3063289