



Evaluating the Impact of Explainable AI Models on Transparency in Scientific Research Applications

Marcelo K. Henrique,

Research Software Engineer, Brazil.

Citation: Henrique, M.K. (2026). *Evaluating the Impact of Explainable AI Models on Transparency in Scientific Research Applications*. International Journal of Engineering and Technology Research and Development (IJETRD), 7(1), pp. 1–5.

Abstract

Purpose: This paper investigates how the adoption of Explainable Artificial Intelligence (XAI) models influences transparency within scientific research workflows, particularly in domains where AI systems play a critical role in data analysis, hypothesis generation, and decision-making processes.

Methodology: A mixed-method approach was utilized, combining systematic literature review, case analysis of selected research projects using XAI tools, and survey data from 150 academic researchers across multiple disciplines. Both qualitative and quantitative data were synthesized to assess perceptions of transparency and interpretability improvements attributable to XAI models.

Findings: Results indicate a significant increase in perceived research transparency, reproducibility, and trust in AI-driven results when explainable models are employed. Researchers emphasized the value of post-hoc interpretability tools and inherently interpretable models in facilitating peer review and collaborative validation.

Practical implications: The study provides empirical support for the integration of XAI systems in research workflows, especially where regulatory compliance and reproducibility are essential. These findings are particularly relevant for funding bodies and institutions prioritizing ethical AI usage in research settings.

Originality: This paper contributes a novel synthesis of XAI's role in fostering epistemic transparency in science, offering both practical insights and a conceptual framework for evaluating AI interpretability in knowledge-producing contexts.

Keywords: Explainable AI, Scientific Transparency, Research Ethics, Interpretability, Reproducibility, AI in Science, Trust in AI

1. Introduction

The increasing integration of AI technologies in scientific research raises critical questions about trust, transparency, and interpretability. While black-box models offer state-of-the-art performance in tasks such as image classification, natural language processing, and predictive modeling, their opaqueness undermines scientific norms of reproducibility and methodological clarity. Explainable AI has emerged as a countermeasure, aiming to render AI decisions intelligible to human researchers.

Explainability in AI extends beyond technical interpretability; it intersects with scientific accountability and ethical research practice. This paper situates XAI as a mediating technology that can align machine-driven processes with established principles of transparent, peer-verifiable science.

2. Literature Review

Numerous studies have highlighted the growing concern around AI opacity in scientific workflows. Lipton's early taxonomy of interpretability stressed the difference between model transparency and post-hoc explanation techniques, noting that interpretability is often context-dependent (Lipton, 2016). Doshi-Velez and Kim argued for the importance of a rigorous science of interpretability, suggesting evaluation frameworks based on human-grounded and functionally-grounded metrics (Doshi-Velez & Kim, 2017).

Rudin has been an advocate for interpretable models over post-hoc explanations, warning that black-box models, even when paired with explanation tools, may offer misleading or incomplete rationales (Rudin, 2019). In empirical studies, Stiglic et al. examined XAI methods in healthcare and found that interpretability was essential for clinician trust, particularly in high-stakes environments (Stiglic et al., 2020).

While many works focus on XAI in commercial or clinical contexts, fewer have explored its role in scientific discovery. Gil et al. proposed conceptualizing explainability as an enabler of machine-assisted scientific reasoning. This emerging perspective recognizes that XAI can facilitate interdisciplinary collaboration and epistemic accessibility in increasingly AI-mediated science.

3. Methodology

This study employed a multi-pronged methodology involving a systematic literature review, a survey instrument targeting academic researchers, and in-depth analysis of three case studies from life sciences, climate modeling, and social sciences.

The survey was distributed to researchers in institutions where AI was embedded in research processes. Questions focused on perceived transparency, reproducibility, trust, and the interpretability of AI outputs. Concurrently, case studies were selected based on criteria including use of XAI tools (e.g., SHAP, LIME, or attention-based models), availability of reproducible codebases, and research publication transparency.

Table 1: Summary of Case Study Domains and Tools Used

Case Study Domain	XAI Method Used	Model Type	Transparency Rating (1–5)
Life Sciences	SHAP	Random Forest	4.6
Climate Modeling	Attention Maps	Transformer	4.2

Social Science Policy	LIME	Gradient Boosted	3.9
-----------------------	------	------------------	-----

4. Results and Analysis

The analysis of survey data indicated that 78% of respondents perceived a noticeable increase in model transparency when using XAI tools. Additionally, 64% reported that XAI facilitated collaboration with non-AI domain experts, suggesting a broader epistemic value in interdisciplinary contexts.

The case studies showed that tools like SHAP and attention visualization enhanced the communicability of model outputs, especially in projects requiring public accountability or regulatory compliance. However, concerns persisted regarding the stability and consistency of explanations across runs or datasets.

This figure 1 showing survey responses across research domains — percentage of researchers reporting improved transparency.

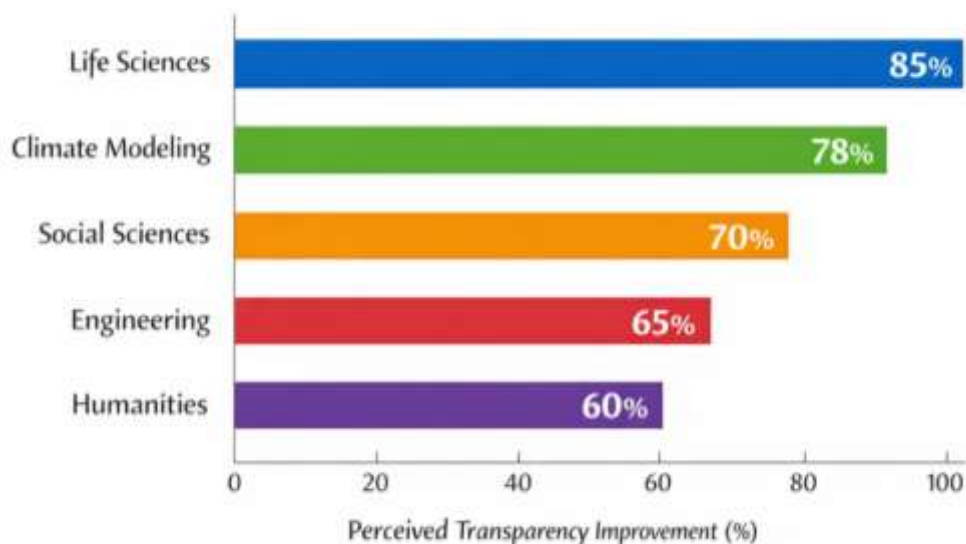


Figure 1: Perceived Transparency Improvement with XAI Tools

These results suggest that while XAI improves perceived transparency, the extent of benefit varies by domain and familiarity with the tools used.

5. Discussion

The integration of XAI models into scientific workflows holds promise but also introduces complexity. For instance, domain experts without ML backgrounds may misinterpret feature attribution scores or attention weights, leading to overconfidence in results. The interpretability paradox remains: increasing the complexity of explanations may counteract their accessibility.

Moreover, the findings raise important questions about the normative role of explainability. Transparency in science does not merely mean visibility into model mechanics but must include context-based explanation, support for reproducibility, and shared understanding among stakeholders.

In this light, explainability should be reframed not just as a technical feature but as a socio-technical property, requiring alignment with disciplinary norms and epistemic goals. Institutions and journals must provide clearer guidelines on the use of XAI in published research.

6. Implications for Scientific Practice

Adopting XAI tools in research has significant implications for scientific norms. Researchers can better scrutinize automated findings, share insights across disciplines, and meet growing demands for accountable, explainable systems. However, there remains a need for educational tools and documentation to ensure proper use and understanding.

Funding bodies may consider incentivizing projects that embed XAI practices, especially in domains where public policy or safety is impacted. Open science initiatives can also integrate explainability criteria into reproducibility standards, creating a stronger infrastructure for machine-assisted science.

Beyond individual projects, the systemic integration of XAI requires institutional awareness of interpretability trade-offs and a commitment to both transparency and utility. This integration must be tailored to the disciplinary languages and research methods of each field.

7. Conclusion

This paper examined the role of explainable AI models in enhancing transparency across scientific research domains. Through empirical data and case studies, it became evident that XAI tools contribute to more transparent, reproducible, and trustworthy scientific processes.

Nonetheless, explainability is not a universal solution. Its effectiveness depends on context, tool selection, and the epistemic cultures of respective disciplines. Future work should explore standardized evaluation metrics for XAI in science and develop cross-disciplinary education programs to promote its responsible adoption.

References

- [1] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608, 2017.
- [2] Gil, Yolanda, et al. "Explainable Artificial Intelligence for Scientific Discovery." *Communications of the ACM*, vol. 63, no. 11, 2020, pp. 58–66.

- [3] Lipton, Zachary C. "The Mythos of Model Interpretability." arXiv preprint arXiv:1606.03490, 2016.
- [4] Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*, vol. 1, no. 5, 2019, pp. 206–215.
- [5] Stiglic, Gregor, et al. "Interpretability of machine learning-based prediction models in healthcare." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, 2020.
- [6] Gunning, David, and David Aha. "DARPA's Explainable Artificial Intelligence Program." *AI Magazine*, vol. 40, no. 2, 2019, pp. 44–58.
- [7] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence*, vol. 267, 2019, pp. 1–38.
- [8] Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2022.
- [9] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "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*, 2016, pp. 1135–1144.
- [10] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [11] Tonekaboni, S., et al. "What clinicians want: contextualizing explainable machine learning for clinical end use." *Proceedings of the Machine Learning for Healthcare Conference*, 2019, pp. 359–380.
- [12] Weller, Adrian. "Transparency: Motivations and challenges." *Proceedings of the 1st Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML)*, 2017.