MANTECH Publications

Journal of Ethical and Explainable AI Systems
Volume 1, Issue 3, September-December 2025

Evaluating Explainability Metrics: Ethical Implications of What We Choose to Explain

Dr. M. Ramesh
Associate Professor
Department of Computer Science and Engineering
Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India
Email: *ramesh.cse@bannariamman.edu.in*

Ms. A. Tanushree Dutta
Assistant Professor
Department of Information Technology
RCC Institute of Information Technology (Rural Campus), Kolkata, West Bengal, India
Email: *tanushreedutta.it25@gmail.com*

Abstract

Explainable Artificial Intelligence (XAI) has emerged as a critical component in ethical and responsible AI deployment, allowing humans to interpret, trust, and audit algorithmic decisions. However, the selection of which aspects of AI systems to explain—features, models, outcomes, or decision paths—has significant ethical implications. Metrics for explainability often focus on technical comprehensibility or accuracy, while overlooking fairness, social context, and stakeholder relevance. This paper examines the ethical dimensions of explainability metrics, analyzing how choices in explanation design influence transparency, bias detection, accountability, and public trust. A conceptual framework is proposed for ethically aligned evaluation of XAI metrics, highlighting trade-offs between interpretability, privacy, and operational effectiveness. Understanding the ethical consequences of metric selection is essential to ensure AI systems serve societal and human-centered goals.

Keywords: *Explainable AI, Ethics, Metrics, Transparency, Accountability, Bias*

INTRODUCTION

Explainable AI allows stakeholders to understand the rationale behind machine learning models and automated decisions. As AI systems influence high-stakes decisions in finance, healthcare, governance, and security, ethical concerns surrounding transparency, accountability, and bias are increasingly prominent.

While technical research often emphasizes methods for generating explanations, the evaluation of these explanations is guided by **explainability metrics**. Metrics may measure fidelity, completeness, consistency, or human interpretability. Yet, the **choice of what to explain and how to measure it** has deep ethical consequences. Excluding relevant factors or focusing on certain features over others may obscure bias, mislead stakeholders, or create false trust.

This paper examines the ethical implications of explainability metric selection and proposes guidelines for evaluating XAI systems with ethical rigor.

EXPLAINABILITY METRICS OVERVIEW

2.1 Fidelity-Based Metrics

- Measure how well an explanation approximates the original model.
- Ethical concern: High fidelity may favor complex models that are technically accurate but incomprehensible to end-users.

2.2 Human-Centered Metrics

- Assess how understandable an explanation is to humans through surveys or task performance.
- Ethical concern: Different stakeholder groups may require different levels or types of explanation, introducing inequities if metrics focus on only one user group.

2.3 Completeness Metrics

- Evaluate whether explanations capture all relevant factors influencing decisions.
- Ethical concern: Omitting sensitive or ethically significant features can lead to unfair outcomes or hidden bias.

2.4 Consistency Metrics

- Measure stability of explanations across similar inputs.
- Ethical concern: Inconsistent explanations can erode trust and accountability.

ETHICAL IMPLICATIONS OF METRIC CHOICES

3.1 Transparency vs. Privacy

Explaining sensitive features (e.g., race, gender) may improve fairness assessment but could violate privacy.

3.2 Bias Visibility

Metrics focusing solely on model fidelity may obscure discriminatory patterns present in input data.

3.3 Stakeholder Relevance

Metrics must consider the needs of different stakeholders: end-users, regulators, auditors, and developers.

Table 1: Ethical Dimensions of Explainability Metrics

Metric Type	Ethical Consideration	Potential Risk
Fidelity	Accurate representation	May obscure interpretability
Human-Centered	Stakeholder comprehension	Excludes minority perspectives
Completeness	Feature coverage	Omits sensitive or critical variables
Consistency	Stability across inputs	Misleads trust if unstable

4. Selecting What to Explain

4.1 Feature-Level Explanations

- Explain contributions of individual input features.
- Ethical concern: May expose private attributes or inadvertently reinforce stereotypes.

4.2 Model-Level Explanations

- Explain how the overall model functions.
- Ethical concern: Complexity may prevent non-technical stakeholders from understanding outcomes.

4.3 Outcome-Level Explanations

- Explain why a specific decision was made.
- Ethical concern: Overemphasis on single decisions may ignore systemic biases.

[Features] ----> Privacy vs Transparency
[Model] ----> Fidelity vs Comprehensibility
[Outcome] ----> Specificity vs Systemic Awareness

Figure 1: Ethical Trade-Offs in Explanation Selection

FRAMEWORK FOR ETHICALLY ALIGNED EXPLAINABILITY EVALUATION

- **Stakeholder Mapping:** Identify all groups affected by AI decisions.
- **Ethical Prioritization:** Determine which ethical principles (fairness, privacy, accountability) are most relevant.
- **Metric Selection:** Choose metrics that balance interpretability, fidelity, and ethical relevance.
- **Evaluation and Iteration:** Continuously assess explanations against real-world impact and ethical standards.

DOMAIN EXAMPLES

6.1 Finance

- Explanation metrics must balance fidelity with fairness, e.g., credit scoring algorithms should not obscure discriminatory patterns.

6.2 Healthcare

- Outcome explanations must be interpretable by clinicians without revealing patient-identifiable data.

6.3 Public Policy

- Explanations for predictive policing or welfare eligibility must be accessible to the public and regulators, emphasizing transparency and accountability.

CHALLENGES IN ETHICAL EVALUATION

- Diverse stakeholder needs create trade-offs in metric design.
- Technical metrics may not align with societal ethical standards.
- Explaining sensitive or legally protected features can create privacy risks.
- Standardization of ethical XAI evaluation remains limited.

FUTURE DIRECTIONS

- Develop multi-dimensional evaluation frameworks incorporating ethics, usability, and fidelity.
- Conduct longitudinal studies on trust and ethical perception linked to different explainability metrics.
- Integrate participatory governance for metric design, ensuring inclusion of minority and marginalized perspectives.
- Explore cross-cultural evaluation of XAI metrics to align explanations with global ethical norms.

CONCLUSION

Explainability metrics in AI do not merely measure technical performance; they have profound ethical implications. What we choose to explain, how we measure interpretability, and whose understanding is prioritized can significantly impact fairness, accountability, privacy, and public trust. This paper highlights the need for ethically informed metric selection and proposes a framework for evaluating XAI systems with human-centered, socially responsible, and context-aware principles. By carefully considering ethical dimensions, AI developers and policymakers can ensure that explainable AI fulfills its promise of transparency, trustworthiness, and societal benefit.

REFERENCES

1. Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence. *Information Fusion*, 58, pp. 82–115.
2. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), pp. 36–43.
3. Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable ML. *arXiv*, pp. 1–13.

4. Molnar, C. (2022). *Interpretable Machine Learning*. Leanpub, pp. 211–258.
5. Hoffman, R. R., et al. (2018). Metrics for explainable AI. *DARPA XAI Program*, pp. 1–27.
6. Ribeiro, M. T., Singh, S., Guestrin, C. (2016). Why should I trust you? *KDD Proceedings*, pp. 1135–1144.
7. Mittelstadt, B. D., et al. (2016). The ethics of algorithms. *Big Data & Society*, 3(2), pp. 1–21.
8. Shneiderman, B. (2020). Human-centered AI. *International Journal of Human–Computer Interaction*, 36(6), pp. 495–504.
9. Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), pp. 1–42.