International Journal on Artificial Intelligence Tools
(2026) 2602002 (9 pages)

DOI: 10.1142/S0218213026020021

World Scientific
www.worldscientific.com

Explainable, Fair, and Trustworthy AI: Current Research, Regulatory Developments, and Future Directions

Sheikh Rabiul Islam [*,¶], Ingrid Russell [†,‖], Douglas Talbert [‡,**] and Md Golam Moula Mehedi Hasan [§,††]

[*]*University at Albany, SUNY, 1400 Washington Avenue, Albany, New York, USA*
[†]*University of Hartford, 200 Bloomfield Avenue, West Hartford, CT, USA*
[‡]*Tennessee Tech University, 1 William L Jones Dr, Cookeville, TN 38505, USA*
[§]*Iona University, 715 North Ave, New Rochelle, NY 10804, USA*
[¶]*sislam7@albany.edu*
[‖]*irussell@hartford.edu*
[**]*dtalbert@tntech.edu*
[††]*mmehedihasan@iona.edu*

Published

The rapid deployment of Artificial Intelligence (AI) in high-stakes domains necessitates robust approaches to transparency, fairness, and trustworthiness. Current advancements in AI performance are outpacing our understanding and ability to govern these systems. This special issue presents research addressing explainability, fairness, and trust as interconnected socio-technical challenges. Accepted papers demonstrate novel techniques for revealing hidden model dependencies, aligning explanations with domain expertise, rigorously benchmarking model classes for explanation robustness, and refining methods for measuring interpretability. We synthesize these contributions, situate them within current policy and standardization. (EU AI Act;[1] NIST AI RMF;[2,3] ISO/IEC 23894 and 42001[4,5]), and connect them to emerging evaluation science in XAI (e.g., BEExAI,[9] Saliency-Bench,[10] F-Fidelity[11]). Finally, we outline a forward-looking agenda emphasizing multi-aspect evaluation, context-sensitive trust, and the development of governance-ready AI systems.

Keywords: Explainable AI; fairness; trustworthy AI; artificial intelligence Act; AI risk management.

1. Introduction: Why Explainable, Fair, and Trustworthy AI Matters

Despite rapid advances in Artificial Intelligence (AI) performance, significant challenges remain in ensuring responsible deployment. AI-based black-box models can

[¶]Corresponding author.

encode spurious correlations, amplify historical biases, and exhibit unpredictable behavior under distributional shift, even when conventional metrics indicate success. High-profile failures have underscored that accuracy alone is insufficient for high-stakes applications — particularly when decisions are irreversible or carry significant social consequences. AI systems are now deeply embedded in socially consequential domains such as healthcare, finance, employment, and public services. In parallel with rapid capability gains, governance expectations have crystallized. The European Union's Artificial Intelligence Act establishes risk-based obligations for transparency, human oversight, post-market monitoring, and documentation, including specific provisions for general-purpose AI models.[1] In the United States, the NIST AI Risk Management Framework (AI RMF 1.0) and its Generative AI Profile provide lifecycle guidance that explicitly treats explainability and fairness as dimensions of trustworthiness.[2,3] International standards have also matured: ISO/IEC 23894:2023 offers AI-specific risk management guidance, and ISO/IEC 42001:2023 specifies organizational requirements for an AI management system.[4,5]

Simultaneously, the field is moving from "produce an explanation" to "measure what explanations buy us." New benchmarks and toolkits have accelerated empirical rigor: BEExAI and Saliency-Bench propose standardized, large-scale evaluation for visual explanations, while F-Fidelity targets faithfulness evaluation with improved robustness across modalities.[9–11] Quantus has emerged as a general-purpose library to compare explanation methods across multiple criteria.[6] In NLP and LLMs, recent studies distinguish faithfulness from self-consistency, highlighting gaps between plausible natural-language rationales and true model behavior.[7,8]

Crucially, trust in AI is increasingly recognized as a socio-technical construct, influenced not only by algorithmic properties but also by users, institutions, regulatory contexts, and deployment environments. Explanations, fairness audits, and interpretability measures play a critical role in mediating this trust by enabling scrutiny, contestability, and informed oversight.

2. Explainability, Fairness, and Trust: Interconnected Challenges

Explainability is often treated as an end goal, but it's more accurately viewed as a means to support trust, accountability, and fairness. Explanations can help reveal hidden model behavior, diagnose failure modes, and enable informed human judgment — but only if they are reliable, meaningful, and aligned with domain knowledge.

Bias in AI systems arises from multiple sources, including skewed datasets, proxy variables, model inductive biases, and deployment feedback loops. Crucially, biased behavior may persist even in high-performing models and may only become visible through targeted analysis or explainability techniques. Therefore, explainability serves as a critical diagnostic tool for fairness assessment, yet it doesn't guarantee fairness on its own — highlighting the need for continuous monitoring and mitigation strategies.

Trust, in turn, is highly context-dependent. What constitutes a trustworthy explanation for a clinician differs from what is required by a regulator, developer, or affected individual. Trust is influenced by explanation stability, semantic alignment, robustness, and perceived legitimacy — dimensions that cut across technical and human-centered evaluation.

3. Where the Field is Moving?

Existing explainability approaches broadly fall into two categories: ***post-hoc* explanations** (e.g., SHAP,[13] LIME,[14] saliency maps[15]) and **inherently interpretable models** (e.g., decision trees, generalized additive models (GAMs), explainable boosting machines (EBMs)). While *post-hoc* methods are widely applicable, they may suffer from instability, low fidelity, and user misinterpretation. Inherently interpretable models offer transparency by design, but may trade off predictive power in some settings. Similarly, fairness research has emphasized **bias detection** — through metrics and audits — more than **bias mitigation**, which remains difficult to generalize across domains and contexts. Moreover, fairness interventions can introduce new trade-offs with accuracy or interpretability.

A persistent limitation across both explainability and fairness research is the lack of standardized evaluation. Many studies rely on user experiments, domain-specific heuristics, or qualitative assessments, making it difficult to compare methods or establish benchmarks. This lack of standardized, objective evaluation frameworks impedes cumulative scientific progress.

Shifting Toward Evaluation Science for XAI. BEExAI and Saliency-Bench propose standardized pipelines and curated datasets for scalable, comparative evaluation of visual explanations, including alignment and causality-based metrics.[9,10] F-Fidelity seeks robustness to out-of-distribution artifacts and information leakage when scoring explanation faithfulness across modalities.[11] The Quantus toolkit offers more than thirty metrics for explanation quality (e.g., faithfulness, robustness, complexity), improving reproducibility and meta-evaluation.[6]

Toward Integrated Governance and Standards. The EU AI Act establishes crucial requirements for high-risk systems, including data governance, technical documentation, transparency, and human oversight — driving demand for explanation, traceability, and post-market monitoring.[1] The NIST AI RMF (2023) and the 2024 Generative AI Profile provide lifecycle guidance (Govern–Map–Measure–Manage) with trustworthiness characteristics such as explainable and interpretable and fair — with harmful bias managed.[2,3] ISO/IEC 23894:2023 and ISO/IEC 42001:2023 translate governance into organizational processes and controls, providing a complementary pathway to demonstrate responsible AI practices.[4,5]

4. Contributions of this Special Issue

4.1. *Quantifying contextual reliance in black-box vision models (Vonderhaar et al.[16])*

Vonderhaar's work focuses on uncovering hidden dependencies in object detection systems by quantifying how model performance varies with the presence or absence of contextual objects. While explainability research in vision has often emphasized saliency maps or localized visual attributions, this contribution shifts attention toward *global behavioral analysis*. By systematically comparing Average Precision across curated test datasets, the approach exposes contextual reliance patterns that would otherwise remain undetected using standard evaluation protocols.

At the same time, this work highlights an open challenge that recurs throughout the special issue: while contextual reliance can be detected, there remains no standardized framework for evaluating *when such reliance is acceptable or harmful* — a question that directly motivates the open challenges discussed in Sec. 5.

4.2. *Domain-aligned explainability in diagnostic machine learning (Hines et al.[17])*

Hines *et al.* address explainability and trust from a domain-specific, human-centered perspective, demonstrating how explainability can be intentionally engineered into diagnostic machine learning systems. Unlike purely *post-hoc* approaches, their pathway to explainability integrates feature engineering and explanation methods in a way that aligns model reasoning with established medical knowledge. This approach underscores the editorial theme that trust arises not from transparency alone, but from *semantic alignment between models and expert understanding*.

Looking ahead, this work also exposes a broader challenge: explainability methods that perform well in one domain may not generalize to others without significant adaptation. The reliance on domain expertise raises questions about scalability, standardization, and evaluation consistency — issues that reappear in Sec. 5 as central obstacles to widespread adoption of trustworthy AI in high-stakes domains.

4.3. *Balancing predictive performance and explanation robustness in tabular data (Lazar et al.[18])*

Lazar's empirical study provides a critical counterpoint to the prevailing assumption that increasingly complex models necessarily yield superior real-world performance. By systematically comparing gradient boosting models and neural architectures across multiple datasets, the paper demonstrates that tree-based methods not only achieve strong predictive accuracy but also deliver more stable and consistent explanations.

At the same time, Lazar's findings point to unresolved challenges that extend beyond model choice. Explanation robustness remains difficult to quantify and compare across methods, datasets, and domains. This limitation highlights the need for standardized evaluation criteria — a challenge that is directly addressed in Goel's conceptual analysis and expanded upon in Sec. 5.

4.4. *Measuring interpretability and explainability (Goel et al.[19])*

A conceptual analysis clarifies distinctions between **model-inherent interpretability** and ***post-hoc* explanations**, identifying the **fragmentation** of current evaluation practices. The paper motivates **standardized, multi-aspect** measures that tie explanation properties (faithfulness, stability, alignment) to downstream tasks like auditing, monitoring, and regulatory reporting — precisely where current policy and standards are heading.[2–5] This provides a foundational framework for the practical implementation outlined in Sec. 6, connecting theoretical understanding with actionable guidelines for evaluation and governance.

5. Open Challenges and Future Directions

5.1. *Standardizing interpretability and explainability evaluation*

Goel's systematic review makes clear that interpretability is currently evaluated through fragmented, aspect-specific proxies rather than unified or comprehensive measures. Lazar's empirical comparison further reinforces this gap by demonstrating the absence of standardized criteria for assessing the robustness of explanations across different models. Similarly, Vonderhaar's analysis of contextual reliance raises additional concerns about how to benchmark acceptable versus harmful dependencies. Collectively, these works underscore the urgent need for standardized, multi-faceted evaluation frameworks that can capture robustness, faithfulness, context sensitivity, and user relevance in a unified way. Recent tools and resources begin to provide building blocks toward such standardization. Quantus offers multi-metric evaluation for explanation methods, while BEExAI and Saliency-Bench supply structured benchmarks for visual explanations. F-Fidelity contributes metrics for out-of-distribution — aware faithfulness assessment.[6–11] Together, these efforts represent a significant momentum toward establishing consistent evaluation standards for XAI.

5.2. *Explainability across domains and contexts*

Hines *et al.* demonstrate that explainability in medical diagnostics requires a tight alignment with domain expertise to foster clinical trust and safe decision-making. In contrast, Vonderhaar shows that contextual effects in vision models can vary significantly across deployment environments, indicating that explanations may not generalize reliably across settings. Together, these findings illuminate a key tension

between domain-specific explainability — tailored to specialized expert needs — and general-purpose methods designed for broad applicability. This tension raises challenges for scalability, cross-domain transferability, and the consistent interpretation of model behavior. At the same time, governance frameworks already mandate transparency, risk controls, and human oversight regardless of domain, highlighting the need for common documentation baselines that can coexist with domain-customized interpretability approaches.[1,2]

5.3. *Balancing performance, robustness, and transparency*

Lazar's results suggest that simpler, tree-based models often outperform deep models in the robustness of their explanations without incurring an accuracy penalty. Complementing this, Hines *et al.* show that explainability can be substantially improved through careful data curation and feature design rather than replacing the underlying model. Together, these findings point to an unresolved challenge: identifying principled trade-offs — and potential synergies — between model performance and transparency, especially when explainability enhancements can emerge from multiple layers of the pipeline. Crucially, these insights advocate for a shift away from solely prioritizing performance, acknowledging that robust and reliable explanations can often be achieved through more modest models and thoughtful design choices.[12]

5.4. *Fairness, hidden stratification, and contextual bias*

Vonderhaar's method reveals contextual dependencies that may align with socioeconomic or environmental factors, while Lazar's benchmarking highlights explanation instability under class imbalance and dataset heterogeneity. Together, these findings point to fairness risks that cannot be detected through aggregate performance metrics alone, underscoring the importance of integrating explainability-driven analyses into fairness assessments. Post-market monitoring workflows anticipated by emerging regulatory guidelines[1] and incorporated into risk registers under ISO/NIST governance frameworks[2,4] help ensure that fairness and transparency considerations extend beyond model development into real-world deployment and ongoing oversight. This underscores the necessity of moving beyond superficial explanations and implementing continuous monitoring for hidden biases and contextual dependencies — a crucial step toward genuinely equitable AI systems.

5.5. *Regulatory, legal, and deployment considerations*

Regulatory frameworks increasingly mandate transparency, accountability, and demonstrable risk mitigation, yet Goel shows that interpretability still lacks objective and standardized baselines. Hines *et al.* emphasize that trust — particularly in clinical contexts — is contingent on explanations that are comprehensible and

aligned with practitioner workflows, while Lazar underscores the need for reliability in high-stakes applications, where explanation instability can undermine safety. Together, these perspectives reveal a widening gap between regulatory expectations and current technical capabilities in interpretability and explainability.

The EU AI Act outlines requirements for transparency, technical documentation, and ongoing monitoring for high-risk systems,[1] while the NIST AI Risk Management Framework and ISO/IEC 42001 provide structured process scaffolding for defining roles, establishing controls, and supporting continuous improvement.[2,5] These frameworks collectively signal a fundamental shift: interpretability is no longer simply a technical pursuit, but a core governance requirement, demanding a holistic approach to risk management and accountability.

5.6. *Participatory and inclusive AI design*

Hines *et al.*'s focus on clinician-aligned explanations illustrates the value of stakeholder-centered design, demonstrating how domain expertise shapes what constitutes a meaningful and trustworthy explanation. Meanwhile, Goel's conceptual separation of interpretability and explainability clarifies where human factors should — and should not — enter the evaluation process, helping distinguish between model-internal assessments and user-facing explanation requirements. Together, these perspectives highlight an ongoing challenge: scaling participatory and stakeholder-informed approaches without compromising methodological rigor. User-centered evaluation frameworks in XAI emphasize empirical, task-grounded assessment of explanations,[8] offering a structured pathway for integrating stakeholder perspectives while maintaining scientific validity. Ultimately, truly effective AI design necessitates a collaborative process, integrating domain knowledge and user insights to ensure that explanations are not only technically sound but also meaningfully understood and trusted.

6. A Pragmatic Roadmap for Researchers and Practitioners

Recognizing the growing imperative for transparent and trustworthy AI, this section provides a brief, clear, and actionable guide for researchers and practitioners seeking to integrate explainability into their work.

- **Select model families** with explanation stability in mind; favor inherently interpretable or explanation-stable model classes where feasible.[14]
- **Design for explainability** from the outset and document assumptions in ways that align with established governance frameworks.[1,2,5]
- **Adopt multi-aspect evaluation** practices using tools such as Quantus, BEExAI, Saliency-Bench, and F-Fidelity.[6,9–11]
- **Institutionalize governance** processes by mapping evaluation artifacts and decisions to the NIST AI RMF and ISO/IEC 23894/42001 standards.[2,4,5]

- **Monitor deployed systems** in real-world conditions with explicit thresholds for contextual reliance and subgroup drift, consistent with post-market monitoring expectations under the EU AI Act.[1]

Ultimately, this roadmap outlines a practical, layered approach to developing and deploying trustworthy AI, prioritizing proactive design, robust evaluation, and ongoing governance.

7. Conclusion

The papers in this special issue demonstrate that explainable, fair, and trustworthy AI emerges from integrated methodological design, rigorous evaluation, and governance-ready practices, not from any single technique. As policy frameworks and technical standards increasingly converge on transparency and risk management, the research frontier is shifting toward measurable, context-appropriate forms of trust: explanations that are faithful, stable, and aligned with human reasoning; fairness audits capable of detecting hidden stratification; and end-to-end pipelines that can withstand regulatory, clinical, and societal scrutiny. Moving forward, sustained collaboration among machine learning researchers, domain experts, social scientists, and policymakers will be essential to ensure that AI systems not only inspire confidence but genuinely earn and deserve human trust — a trust built on a foundation of transparency, accountability, and demonstrable impact.

ORCID

Sheikh Rabiul Islam https://orcid.org/0000-0001-9610-0230

Ingrid Russell https://orcid.org/0000-0002-1328-3009

Douglas Talbert https://orcid.org/0000-0001-8073-1134

Md Golam Moula Mehedi Hasan https://orcid.org/0000-0001-7309-7200

References

1. European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). EUR-Lex (2024), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
2. National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0) (NIST AI 100-1) (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
3. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative AI Profile (NIST-AI-600-1) (2024), https://www.nist.gov/itl/ai-risk-management-framework.
4. ISO/IEC, ISO/IEC 23894:2023 — Information technology — Artificial intelligence — Guidance on risk management (2023a), https://www.iso.org/standard/77304.html.
5. ISO/IEC, ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system (2023b), https://www.iso.org/standard/42001.

6. A. Hedström *et al.*., Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond, *J. Mach. Learn. Res.* **24**(34) (2023) 1–11, http://jmlr.org/papers/v24/22-0142.html.
7. L. Parcalabescu and A. Frank, On measuring faithfulness or self-consistency of natural language explanations, in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*) (2024), pp. 6048–6089.
8. Q. Lyu, M. Apidianaki and C. Callison-Burch, Towards faithful model explanation in NLP: A survey, *Comput. Linguist.* **50**(2) (2024) 657–723, doi:10.1162/coli_a_00511.
9. S. Sithakoul, S. Meftah and C. Feutry, BEE*X*AI: Benchmark to evaluate explainable AI, in *World Conf. Explainable Artificial Intelligence* (Springer Nature, Cham, Switzerland, 2024), pp. 445–468.
10. Y. Zhang, J. Song, S. Gu, T. Jiang, B. Pan, G. Bai and L. Zhao, Saliency-bench: A comprehensive benchmark for evaluating visual explanations, in *Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining Volume 2* (ACM, 2025), pp. 5924–5935.
11. X. Zheng, F. Shirani, Z. Chen, C. Lin, W. Cheng, W. Guo and D. Luo, F-Fidelity: A robust framework for faithfulness evaluation of explainable AI, *The Thirteenth International Conference on Learning Representations* arXiv:2410.02970 (2024).
12. C. Rudin, (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* **1**(5) (2019) 206–215.
13. E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher and G. Groh, SHAP-based explanation methods: A review for NLP interpretability, in *Proc. 29th Int. Conf. Computational Linguistics* (2022), pp. 4593–4603.
14. M. T. Ribeiro and S. Singh and C. Guestrin, Model-agnostic interpretability of machine learning, arXiv:1606.05386 (2016).
15. C. Etmann, S. Lunz, P. Maass and C. B. Schönlieb, On the connection between adversarial robustness and saliency map interpretability, arXiv:1905.04172 (2019).
16. L. Vonderhaar, T. Elvira and O. Ochoa, Measuring contextual reliance of object detection models: A black box explainability method, Special issue on FLAIRS-37 special track on Explainable, Fair, and Trustworthy AI, *Int. J. Artif. Intell. Tools* (2026).
17. B. Hines, D. Talbert and S. Anton, Explainable diagnostic machine learning models with feature dependence: A case study in smart knee implants, Special Issue on FLAIRS-37 Special Track on Explainable, Fair, and Trustworthy AI, *Int. J. Artif. Intell. Tools* (2026).
18. A. Lazar, P. Pokhrel and S. Das, Beyond accuracy: A comprehensive comparative study of gradient boosting versus tabular deep learning and explainability techniques for mixed-type tabular data models using SHAP and LIME, Special Issue on FLAIRS-37 Special Track on Explainable, Fair, and Trustworthy AI, *Int. J. Artif. Intell. Tools* (2026).
19. P. Goel and R. Weber, Measuring interpretability: A systematic literature review of interpretability measures in artificial intelligence, Special Issue on FLAIRS-37 Special Track on Explainable, Fair, and Trustworthy AI, *Int. J. Artif. Intell. Tools* (2026).