Generated by Scopus AI, Mon Nov 10 2025

Research inquiries in artificial intelligence interpretability

Interpreting Artificial Intelligence: Structure, Reasoning, and Research Gaps through Machine Learning Interpretability Methods

Quick Reference

Key Findings Table

Theme	Key Insights	Research Gaps	Citations
Intrinsic vs. Post-hoc Interpretability	Intrinsic methods offer faithful, structured insights; post-hoc methods (e.g., SHAP, LIME) are flexible but risk misleading explanations	Need for hybrid approaches and context-aware selection	[1] [2] [3]
Evaluation Metrics & Standardization	Lack of unified metrics; frameworks like Co-12, XAIB, and EXACT are emerging	Persistent ambiguity between interpretability and explainability; need for domain-agnostic protocols	[4] [5] [6]
Scalability & Emerging Paradigms	Adaptation to federated/self-supervised learning is nascent; computational trade-offs are critical	Scalability, privacy, and integration with new architectures	[7] [8] [9]
User/Domain Knowledge Integration	Embedding domain/user knowledge improves explanation fidelity and relevance	Systematic, scalable integration methods needed	[10] [11] [12]
Visual & Quantitative Explanations	Visual (e.g., Grad-CAM) and concept-based methods enhance local interpretability; quantitative metrics (fidelity, compactness) are widely used	Metrics often miss ethical/contextual dimensions	[13] [14] [15]

Direct Answer

Inquiries into the structure, reasoning, inference, and implications of AI using machine learning interpretability methods focus on revealing internal model behaviors via intrinsic design or post-hoc explanations (e.g., LIME, SHAP, visual tools). These studies assess explanation quality using metrics like fidelity and compactness, and raise concerns about the tradeoff between performance and interpretability, standardization of evaluation, risks of misleading explanations, and scalability in federated/self-supervised or domain-specific systems. Research gaps include theoretical ambiguities (interpretability vs. explainability), practical scalability, and human-centric evaluation frameworks.

Study Scope

- **Time Period:** 2018–2024 (emphasis on recent meta-analyses and empirical studies)

- **Disciplines:** Computer Science, Human-Computer Interaction, Ethics, Healthcare, Finance, Autonomous Systems
- **Methods:** Systematic reviews, empirical benchmarking, user studies, theoretical analysis, case studies in regulated domains

Assumptions & Limitations

- **Assumptions:** Interpretability and explainability are distinct but overlapping; high-stakes domains require both technical and human-centered evaluation; current metrics are necessary but not sufficient for real-world trust.
- **Limitations:** Lack of universal definitions; most studies focus on tabular/image data, less on text or multimodal; limited longitudinal studies on real-world deployment; scalability and privacy in federated/self-supervised learning are underexplored.

Suggested Further Research

- Develop standardized, domain-agnostic evaluation protocols that clearly distinguish interpretability from explainability.
- Create robust quantitative metrics encompassing technical fidelity, ethical/fairness, and human-centered aspects.
- Innovate scalable interpretability methods for federated, self-supervised, and resource-constrained AI.
- Advance interactive, user-centric, and context-aware explanation systems.
- Foster interdisciplinary collaboration to bridge technical, ethical, and societal perspectives.

1. Introduction

Background and Motivation

The rapid proliferation of complex AI models—especially deep neural networks—has led to unprecedented performance in domains ranging from healthcare to finance. However, this progress has intensified the "black-box" problem: the opacity of model decision-making processes, which undermines trust, accountability, and regulatory compliance in high-stakes applications. Interpretability has thus emerged as a critical requirement, not only for technical validation but also for human-centered understanding and ethical deployment. The field of machine learning interpretability has responded with a spectrum of methods, from models designed for transparency (intrinsic) to post-hoc explanation techniques. This report systematically explores how these methods illuminate the structure, reasoning, and implications of AI, and identifies the theoretical and practical concerns that define the current research landscape and its gaps [16] [17] [18].

2. Revealing the Structure of AI Models through Interpretability Methods

Intrinsic vs. Post-hoc Interpretability Approaches

- **Intrinsic Interpretability:** Models are designed for transparency (e.g., decision trees, rule-based systems, interpretable-by-design neural networks). These approaches provide direct, faithful insights into model structure and decision logic, often with negligible computational overhead and without the need for surrogate explanations [1] [19].

- **Post-hoc Interpretability:** Explanations are generated after model training, using techniques like LIME, SHAP, and Grad-CAM. These methods are model-agnostic and flexible, applicable across architectures (CNNs, LSTMs, etc.), but may risk generating misleading or artifact-based explanations, especially if not aligned with domain knowledge [1] [3].

- **Comparative Insights:** Intrinsic methods tend to outperform post-hoc in explanation fidelity and consistency, particularly in regulated domains. However, post-hoc methods remain essential for legacy or high-performing black-box models [1] [3].

Quantitative Evaluation of Explanations

- **Metrics:** Fidelity (how well explanations reflect model behavior), correctness (alignment with ground truth or expert reasoning), compactness, and cognitive load are key properties [4] [20].

- **Frameworks:** The Co-12 framework and benchmarking platforms like EXACT and XAIB provide multi-dimensional, modular evaluation environments. However, many popular methods struggle to outperform random baselines in explanation fidelity, highlighting the need for more robust metrics [20] [21].

- **Tools:** SHAP and LIME are widely used for local/global interpretability, with SHAP often providing better alignment with expert reasoning in domains like healthcare [22].

Integrating Domain and User Knowledge

- **Systematic Integration:** Embedding domain knowledge can be achieved by incorporating compatibility terms in cost functions (e.g., KICE for counterfactuals), expert-guided data structuring, and domain-aware synthetic neighborhoods [10] [23].

- **Benefits:** Improves explanation fidelity, plausibility, and user trust, especially in sensitive domains [24].

- **Challenges:** Scalability and automation of domain knowledge integration remain open problems.

Computational Trade-offs in Regulated Domains

- **Pre hoc/Co hoc vs. Post-hoc:** Integrating interpretability during model training (pre hoc/co hoc) yields more faithful, consistent explanations with lower computational cost and can even improve accuracy (up to 3% in some regulated domains) [19] [25].

- **Hybrid Approaches:** Combining interpretable models with black-boxes and explainability tools (e.g., logistic regression + neural networks + SHAP/LIME) balances accuracy and interpretability [25].

- **Trade-offs:** Post-hoc methods incur higher computational costs and may produce variable fidelity, which is problematic in real-time or resource-constrained settings [26] [27].

Synthesis: Interpretability methods reveal AI model structure by clarifying feature contributions and decision logic. Intrinsic approaches offer more reliable insights, while post-hoc methods provide flexibility but require careful evaluation and, ideally, domain knowledge integration. Quantitative metrics and benchmarking frameworks are advancing, but standardization and scalability remain key challenges.

3. Reasoning and Inference in AI via Interpretability

Complementary Interpretability Methods: SHAP and LIME

- **Local vs. Global Explanations:** LIME provides local, instance-specific explanations; SHAP offers both local and global feature attributions, quantifying contributions across the model [17] [28].
- **Complementarity:** SHAP excels in global interpretability, LIME in clarifying individual predictions. Their combined use is especially valuable in sensitive domains (e.g., healthcare, finance) [29].
- **Stability Enhancements:** Deterministic and stable LIME variants (e.g., ST-LIME, DLIME) improve reliability and fidelity, addressing randomness-induced variability [30] [31].

Visual Explanation Techniques in Deep Learning

- **Methods:** Grad-CAM, saliency maps, and attention visualizations highlight influential input regions, making deep neural network inference more transparent [13] [32].
- **User Understanding:** Visual explanations, especially when combined with semantic or class prototypes, improve user comprehension and trust [33] [34].
- **Robustness:** Quantitative metrics (e.g., Decisive Saliency Maps) help users assess explanation reliability and potential risks [35].

Stability and Reliability of Local Explanations

- **Stable LIME Variants:** Approaches like ST-LIME, DLIME, and Bayesian-enhanced LIME (BE-LIME) significantly improve explanation stability and fidelity across data modalities (tabular, time series, images) [9] [31] [36].
- **Limitations:** Even with stability improvements, LIME variants may struggle to identify causative features in complex domains (e.g., audio), indicating that stability does not guarantee veracity [37] [38].

Combining Visual and Concept-based Explanations

- **ConceptLIME and CAVLI:** These frameworks integrate visual and concept-based explanations, providing human-understandable concepts and quantifying their influence on model decisions [39] [40].
- **Interactive Analytics:** Tools like ConceptExplainer and immersive 3D visualizations enable users to explore concept spaces and explanation granularity, enhancing comprehension and trust [41] [42].
- **User Studies:** Multimodal explanations (visual + textual) and interactive interfaces (e.g., SHAPRap) facilitate better understanding and confidence [14].

Synthesis: Interpretability techniques elucidate AI reasoning by making decision paths transparent, supporting both technical and non-technical users. The interplay of local/global, visual/conceptual, and stable explanation methods is central to advancing user trust and practical deployment.

4. Theoretical and Practical Concerns in AI Interpretability

Definitional Ambiguity: Interpretability vs. Explainability

- **Conceptual Ambiguity:** Interpretability (intrinsic model transparency) and explainability (post-hoc, human-centered explanations) are often conflated, complicating method selection and evaluation [18] [43] [44].
- **Frameworks:** Recent efforts (e.g., XAIB, EXACT, Co-12) aim to clarify these distinctions and provide modular, multi-dimensional evaluation protocols [4] [44].
- **Impact:** Ambiguity hinders standardization and consensus on evaluation criteria, leading to fragmented and anecdotal practices [4] [45].

Balancing Accuracy and Interpretability in High-Stakes Domains

- **Trade-off Strategies:** Simulatability operation count (SOC), additive tree models, and hybrid architectures (e.g., TRACER) enable systematic balancing of accuracy and interpretability [46] [47] [48].
- **Domain Examples:** In healthcare and criminal justice, interpretable models can match or outperform black-boxes in both accuracy and trustworthiness [49] [50].
- **Fairness:** Fairness-aware interpretable modeling (FAIM) improves fairness without sacrificing accuracy, supporting ethical deployment [51].

Disentangling Computational and Human-Centered Evaluation

- **Methodologies:** Frameworks like EXSUM, GEF, and user-centered evaluation protocols (e.g., GWAP, HIVE) enable separate assessment of computational faithfulness and human intelligibility [52] [53] [54].
- **Classification:** Revised taxonomies distinguish intrinsic (interpretability) from extrinsic (explainability), supporting targeted evaluation [44].
- **Benchmarks:** XAIB and similar platforms provide modular, extensible environments for computational evaluation, complementing human-centered studies [55].

Standardization through Ontology-Based and Modular Benchmarks

- **Platforms:** IHRAS, cp3-bench, and scenario-based frameworks enable standardized, transparent, and scalable evaluation across diverse models and datasets [56] [57] [58].
- **Limitations:** Domain-specific considerations remain critical; standardized benchmarks may not always predict real-world performance [58].

Synthesis: Theoretical and practical concerns center on definitional clarity, standardization, and the balance between technical and human-centered evaluation. Modular, ontology-based frameworks are promising but require further refinement and domain adaptation.

5. Limitations, Challenges, and Research Gaps

Current Limitations and Challenges

- **Transparency & Comprehensibility:** Many models remain opaque, and explanations may be incomplete or misleading, especially in complex or high-stakes domains [59] [60] [61].
- **Standardized Metrics:** Absence of uniform, objective metrics hinders rigorous assessment and comparison of interpretability methods [62] [63] [64].
- **Cognitive Bias & Deceptive Transparency:** Explanations can mislead users, necessitating frameworks to ensure fidelity and authenticity [65].
- **Domain Expertise & Data:** Lack of interpretable datasets and domain-specific expertise limits effective explainability, particularly in healthcare and finance [62] [66] [67].
- **Scalability:** Integration with federated/self-supervised learning and deployment in resource-constrained environments (e.g., edge AI) are underdeveloped [61] [67].

Standardized Metrics and Protocols for Assessment

- **Existing Tools:** SHAP, LIME, Co-12, and human factors heuristic analysis are widely used, but often lack domain-agnostic applicability [4] [22] [68].
- **Proposed Frameworks:** MM4XAI-AE, integrative XAI frameworks, and normalized multi-dimensional evaluation scales are emerging [68] [69].
- **Challenges:** Current protocols often fail to address the full spectrum of technical, ethical, and human-centered needs.

Scalability and Integration with Emerging AI Paradigms

- **Federated Learning:** Methods like FL-IRT, FedCBM, and FedFTL-R improve interpretability and scalability while preserving privacy [70] [71] [72].
- **Self-supervised Learning:** Unified training approaches (e.g., OFA) and accelerated model-agnostic explanations (AcME) offer scalable, interpretable models [7] [8].
- **Blockchain Integration:** Enhances privacy, security, and interpretability in sensitive domains [73].

Limitations of Quantitative Metrics in Real-World Applications

- **Ethical & Contextual Gaps:** Metrics like Co-12 often miss fairness, robustness, and contextual factors critical in high-stakes applications [74] [75] [76].
- **Data Limitations:** Small, variable datasets undermine reliability of bias and risk assessments [75].
- **Societal Alignment:** Technical metrics may misalign with societal and ethical priorities, necessitating recalibrated frameworks [77] [78].

Adapting Human Factors Evaluation for Resource-Constrained Environments

- **Formalization & Automation:** Protocols must be formalized and tool-supported to reduce reliance on expert availability and improve consistency [79].
- **AI-Assisted Evaluation:** LLMs (e.g., GPT-4o) can assist but require calibration to avoid false positives [80].
- **Cognitive Models:** Integrating human cognitive models and domain-specific heuristics enhances relevance in specialized, resource-limited settings [81] [82].

Identified Research Gaps and Future Directions

- **Standardized, Domain-Agnostic Protocols:** Need for robust, universally applicable evaluation frameworks that distinguish interpretability from explainability [60] [65] [83].
- **Context-Aware, User-Centric Explanations:** Development of interactive, adaptive explanation systems that dynamically incorporate user feedback and domain knowledge [60] [65].
- **Scalability & New Paradigms:** Strategies for integrating interpretability into federated, self-supervised, and edge AI remain underexplored [61] [67].
- **Interdisciplinary Collaboration:** Bridging technical, ethical, and societal perspectives is essential for responsible AI deployment [60] [83].

Synthesis: Despite methodological advances, significant gaps persist in standardization, scalability, and human-centric evaluation. Addressing these requires innovative, interdisciplinary approaches and robust, context-sensitive frameworks.

6. Conclusion

Summary of Key Insights

- **Structure:** Interpretability methods, both intrinsic and post-hoc, are essential for revealing AI model structure, with intrinsic approaches offering more reliable insights but post-hoc methods providing necessary flexibility [16] [17].
- **Reasoning:** Techniques like SHAP, LIME, and visual/concept-based explanations elucidate AI reasoning and inference, supporting both technical validation and user trust [17] [18].
- **Concerns:** Theoretical ambiguities, lack of standardized metrics, cognitive bias risks, and scalability challenges are persistent concerns [18] [60].
- **Research Gaps:** Standardized, domain-agnostic evaluation protocols, context-aware explanations, and scalable methods for emerging AI paradigms are critical areas for future research [60] [65] [83].

Recommendations for Future Research

- **Standardization:** Develop and adopt robust, modular evaluation frameworks that clearly distinguish interpretability from explainability and are applicable across domains and architectures.

- **Human-Centric Design:** Prioritize user-centered, interactive, and context-aware explanation systems that dynamically integrate domain knowledge and user feedback.

- **Scalability:** Innovate interpretability methods suitable for federated, self-supervised, and resource-constrained AI, ensuring privacy and efficiency.

- **Interdisciplinary Collaboration:** Foster partnerships across technical, ethical, and societal domains to ensure responsible, trustworthy AI deployment.

- **Ethical Integration:** Expand quantitative metrics to include fairness, robustness, and societal impact, aligning technical evaluation with real-world needs.

In summary, the landscape of AI interpretability is rapidly evolving, with significant progress in methods and evaluation frameworks. However, persistent challenges in standardization, scalability, and human-centricity highlight the need for continued, interdisciplinary research to ensure AI systems are not only powerful but also transparent, trustworthy, and aligned with societal values.

References

1. ExaM: Unsupervised Concept-Based Representation Learning to Better Explain Models in Vision Tasks Heritier, M., Mekhazni, D., Leblond-Menard, C., (...), Granger, E. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2025 https://www.scopus.com/pages/publications/105017857017?origin=scopusAI

2. Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales Resck, L.E., Raimundo, M.M., Poco, J. Findings of the Association for Computational Linguistics: NAACL 2024 - Findings, 2024 https://www.scopus.com/pages/publications/85197858157?origin=scopusAI

3. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence Vilone, G., Rizzo, L., Longo, L. CEUR Workshop Proceedings, 2020 https://www.scopus.com/pages/publications/85099341984?origin=scopusAI

4. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI Nauta, M., Trienes, J., Pathak, S., (...), Seifert, C. ACM Computing Surveys, 2023 https://www.scopus.com/pages/publications/85168800297?origin=scopusAI

5. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI Shin, D. International Journal of Human Computer Studies, 2021 https://www.scopus.com/pages/publications/85094928986?origin=scopusAI

6. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence Ali, S., Abuhmed, T., El-Sappagh, S., (...), Herrera, F. Information Fusion, 2023 https://www.scopus.com/pages/publications/85159601901?origin=scopusAI

7. One For All: A Unified Approach to Classification and Self-explanation Naouar, M., Vogt, Y., Boedecker, J., (...), Kalweit, M. Lecture Notes in Computer Science, 2026 https://www.scopus.com/pages/publications/105018113682?origin=scopusAI

8. AcME—Accelerated model-agnostic explanations: Fast whitening of the machine-learning black box Dandolo, D., Masiero, C., Carletti, M., (...), Susto, G.A. Expert Systems with Applications, 2023 https://www.scopus.com/pages/publications/85141498082?origin=scopusAI

9. Explainable object detection for aircraft visual landing system based on BE-LIME method Chen, X., Zhang, R., Dong, L., Liu, J. Aerospace Systems, 2025 https://www.scopus.com/pages/publications/105012032366?origin=scopusAI

10. Integrating Prior Knowledge in Post-hoc Explanations Jeyasothy, A., Laugel, T., Lesot, M.-J., (...), Detyniecki, M. Communications in Computer and Information Science, 2022 https://www.scopus.com/pages/publications/85135031206?origin=scopusAI

11. Interpretable AI for bio-medical applications Sathyan, A., Weinberg, A.I., Cohen, K. Complex Engineering Systems, 2022 https://www.scopus.com/pages/publications/85173557681?origin=scopusAI

12. Advancement in Explainable AI: Bringing Transparency and Interpretability to Machine Learning Models for Use in High-Stakes Decisions David, R., Shankar, H., Kura, P., (...), Karkuzhali, S. 2025 International Conference on Emerging Smart Computing and Informatics, ESCI 2025, 2025 https://www.scopus.com/pages/publications/105007282593?origin=scopusAI

13. Attention branch network: Learning of attention mechanism for visual explanation Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019 https://www.scopus.com/pages/publications/85078230819?origin=scopusAI

14. Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives Chromik, M. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021 https://www.scopus.com/pages/publications/85115234045?origin=scopusAI

15. Better metrics for evaluating explainable artificial intelligence Rosenfeld, A. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2021 https://www.scopus.com/pages/publications/85109683014?origin=scopusAI

16. A Review of Explainable Artificial Intelligence Lin, K.-Y., Liu, Y., Li, L., Dou, R. IFIP Advances in Information and Communication Technology, 2021 https://www.scopus.com/pages/publications/85115256359?origin=scopusAI

17. Interpretable Deep Learning for Enhanced AI Trust and Clarity Rumapea, H., Manalu, D.R., Rumapea, Y.Y.P. Journal of Artificial Intelligence and Technology, 2025 https://www.scopus.com/pages/publications/105018214251?origin=scopusAI

18. Assessing interpretation capacity in machine learning: A critical review Haddouchi, M., Berrado, A. ACM International Conference Proceeding Series, 2018 https://www.scopus.com/pages/publications/85062793856?origin=scopusAI

19. Pre Hoc and Co Hoc Explainability: Frameworks for Integrating Interpretability into Machine Learning Training for Enhanced Transparency and Performance Acun, C., Nasraoui, O. Applied Sciences (Switzerland), 2025 https://www.scopus.com/pages/publications/105010304758?origin=scopusAI

20. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations Abdul, A., Von Der Weth, C., Kankanhalli, M., Lim, B.Y. Conference on Human Factors in Computing Systems - Proceedings, 2020 https://www.scopus.com/pages/publications/85091285319?origin=scopusAI

21. EXACT: Towards a platform for empirically benchmarking machine learning model explanation methods Clark, B., Wilming, R., Dox, A., (...), Haufe, S. Measurement: Sensors, 2025 https://www.scopus.com/pages/publications/85214348521?origin=scopusAI

22. Explainable AI for Diabetes Risk Assessment: Enhancing Clinical Decision Support with Interpretable Machine Learning Models Reddy, C.S., Annamalai, M. Proceedings of 5th International Conference on Soft Computing for Security Applications, ICSCSA 2025, 2025 https://www.scopus.com/pages/publications/105018462818?origin=scopusAI

23. Integrating human knowledge for explainable AI Cappuccio, E., Kathirgamanathan, B., Rinzivillo, S., (...), Andrienko, N. Machine Learning, 2025 https://www.scopus.com/pages/publications/105018850418?

origin=scopusAI

24. Post-hoc recommendation explanations through an efficient exploitation of the DBpedia category hierarchy Du, Y., Ranwez, S., Sutton-Charani, N., Ranwez, V. Knowledge-Based Systems, 2022 https://www.scopus.com/pages/publications/85127154002?origin=scopusAI

25. A Hybrid Mathematical Framework Combining Logistic Regression and Neural Networks with Explainable AI Techniques for Mental Health Prediction Humayun, A., Nawi, M.A.B.A., Siddiqui, M.I., (...), Babalola, A. Contemporary Mathematics (Singapore), 2025 https://www.scopus.com/pages/publications/105017053896?origin=scopusAI

26. Towards Explainable Artificial Intelligence in Machine Learning: A study on efficient Perturbation-Based Explanations Gómez-Talal, I., Azizsoltani, M., Bote-Curiel, L., (...), Singh, A. Engineering Applications of Artificial Intelligence, 2025 https://www.scopus.com/pages/publications/105004547922?origin=scopusAI

27. Transaction Scam Detection Using Flask Frame Work Rohini, M., Archana, T., Thulasi Raman, S., Mukilan, R. Proceedings of the International Conference on Multi-Agent Systems for Collaborative Intelligence, ICMSCI 2025, 2025 https://www.scopus.com/pages/publications/105000503510?origin=scopusAI

28. Comparative Evaluation of Post-Hoc Explainability Methods in AI: LIME, SHAP, and Grad-CAM Narkhede, J. 4th International Conference on Sustainable Expert Systems, ICSES 2024 - Proceedings, 2024 https://www.scopus.com/pages/publications/85214838282?origin=scopusAI

29. Explainable Artificial Intelligence for Analytical Customer Relationship Management in Banking and Finance Kumar, S., Kumar, B.A., Ravi, V. Lecture Notes in Networks and Systems, 2024 https://www.scopus.com/pages/publications/85200683596?origin=scopusAI

30. A Model-Agnostic Interpretability Approach Based on Enhanced Hierarchical Clustering Ding, X., Zhu, Y., Yu, Q., Huang, C. Proceedings - 2024 IEEE 24th International Conference on Software Quality, Reliability and Security Companion, QRS-C 2024, 2024 https://www.scopus.com/pages/publications/85209781225?origin=scopusAI

31. Improving Local Interpretable Model-agnostic Explanations Stability Elgezawy, A., Abdul-kader, H., Elsaid, A. International Journal of Intelligent Engineering and Systems, 2024 https://www.scopus.com/pages/publications/85208126010?origin=scopusAI

32. A Dual Approach with Grad-CAM and Layer-Wise Relevance Propagation for CNN Models Explainability Mishra, A., Malhotra, M. Communications in Computer and Information Science, 2025 https://www.scopus.com/pages/publications/85218443701?origin=scopusAI

33. Global and Local Explanations for Skin Cancer Diagnosis Using Prototypes Santiago, C., Correia, M., Verdelho, M.R., (...), Barata, C. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2023 https://www.scopus.com/pages/publications/85180624513?origin=scopusAI

34. Improving interpretability of deep neural networks with semantic information Dong, Y., Su, H., Zhu, J., Zhang, B. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017 https://www.scopus.com/pages/publications/85044463205?origin=scopusAI

35. Visual explanation and robustness assessment optimization of saliency maps for image classification Xu, X., Mo, J. Visual Computer, 2023 https://www.scopus.com/pages/publications/85141775769?origin=scopusAI

36. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability Zafar, M.R., Khan, N. Machine Learning and Knowledge Extraction, 2021 https://www.scopus.com/pages/publications/85117351889?origin=scopusAI

37. ON THE VERACITY OF LOCAL, MODEL-AGNOSTIC EXPLANATIONS IN AUDIO CLASSIFICATION: TARGETED INVESTIGATIONS WITH ADVERSARIAL EXAMPLES Praher, V., Prinz,

K., Flexer, A., Widmer, G. Proceedings of the International Society for Music Information Retrieval Conference, 2021 https://www.scopus.com/pages/publications/85207891542?origin=scopusAI

38. ON THE VERACITY OF LOCAL, MODEL-AGNOSTIC EXPLANATIONS IN AUDIO CLASSIFICATION: TARGETED INVESTIGATIONS WITH ADVERSARIAL EXAMPLES Praher, V., Prinz, K., Flexer, A., Widmer, G. Proceedings of the 22nd International Conference on Music Information Retrieval, ISMIR 2021, 2021 https://www.scopus.com/pages/publications/85123697770?origin=scopusAI

39. A Concept-Based Local Interpretable Model-Agnostic Explanation Approach for Deep Neural Networks in Image Classification Tan, L., Huang, C., Yao, X. IFIP Advances in Information and Communication Technology, 2024 https://www.scopus.com/pages/publications/85190862721?origin=scopusAI

40. CAVLI - Using image associations to produce local concept-based explanations Shukla, P., Bharati, S., Turk, M. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2023 https://www.scopus.com/pages/publications/85170822912?origin=scopusAI

41. ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective Huang, J., Mishra, A., Kwon, B.C., Bryan, C. IEEE Transactions on Visualization and Computer Graphics, 2023 https://www.scopus.com/pages/publications/85139506419?origin=scopusAI

42. Immersive analysis for explainable artificial intelligence: A conceptual approach Montilla-López, J., Ramírez-González, G. 2025 IEEE Colombian Conference on Communications and Computing, COLCOM 2025 - Conference Proceedings, 2025 https://www.scopus.com/pages/publications/105019973317?origin=scopusAI

43. Measuring Interpretability: An Investigation of Domain Independent Interpretability Goel, P., Weber, R.O. 2025 6th International Conference on Pattern Recognition and Machine Learning, PRML 2025, 2025 https://www.scopus.com/pages/publications/105017968494?origin=scopusAI

44. Classifying XAI Methods to Resolve Conceptual Ambiguity Dib, L., Capus, L. Technologies, 2025 https://www.scopus.com/pages/publications/105017496034?origin=scopusAI

45. Notions of explainability and evaluation approaches for explainable artificial intelligence Vilone, G., Longo, L. Information Fusion, 2021 https://www.scopus.com/pages/publications/85107637272?origin=scopusAI

46. A Machine Learning Model Selection considering Tradeoffs between Accuracy and Interpretability Nazir, Z., Kaldykhanov, D., Tolep, K.-K., Park, J.-G. 2021 13th International Conference on Information Technology and Electrical Engineering, ICITEE 2021, 2021 https://www.scopus.com/pages/publications/85123350689?origin=scopusAI

47. Building more accurate decision trees with the additive tree Luna, J.M., Gennatas, E.D., Ungar, L.H., (...), Valdes, G. Proceedings of the National Academy of Sciences of the United States of America, 2019 https://www.scopus.com/pages/publications/85072786403?origin=scopusAI

48. TRACER: A Framework for Facilitating Accurate and Interpretable Analytics for High Stakes Applications Zheng, K., Cai, S., Chua, H.R., (...), Ooi, B.C. Proceedings of the ACM SIGMOD International Conference on Management of Data, 2020 https://www.scopus.com/pages/publications/85086249443?origin=scopusAI

49. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead Rudin, C. Nature Machine Intelligence, 2019 https://www.scopus.com/pages/publications/85069492292?origin=scopusAI

50. EADTC: An Approach to Interpretable and Accurate Crime Prediction Ma, Y., Nakamura, K., Lee, E.-J., Bhattacharyya, S.S. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2022 https://www.scopus.com/pages/publications/85142744241?origin=scopusAI

51. FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare Liu, M., Ning, Y., Ke, Y., (...), Liu, N. Patterns, 2024 https://www.scopus.com/pages/publications/85207495965?origin=scopusAI

52. EXSUM: From Local Explanations to Model Understanding Zhou, Y., Ribeiro, M.T., Shah, J. NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 2022 https://www.scopus.com/pages/publications/85131606367?origin=scopusAI

53. Towards a Unified Framework for Evaluating Explanations Pinto, J.D., Paquette, L. CEUR Workshop Proceedings, 2024 https://www.scopus.com/pages/publications/85210828009?origin=scopusAI

54. Towards a Comprehensive Human-Centred Evaluation Framework for Explainable AI Donoso-Guzmán, I., Ooge, J., Parra, D., Verbert, K. Communications in Computer and Information Science, 2023 https://www.scopus.com/pages/publications/85175966305?origin=scopusAI

55. Open and Extensible Benchmark for Explainable Artificial Intelligence Methods Moiseev, I., Balabaeva, K., Kovalchuk, S. Algorithms, 2025 https://www.scopus.com/pages/publications/85218626845?origin=scopusAI

56. IHRAS: Automated Medical Report Generation from Chest X-Rays via Classification, Segmentation, and LLMs Rodrigues, G.A.P., Serrano, A.L.M., Bispo, G.D., (...), Meneguette, R.I. Bioengineering, 2025 https://www.scopus.com/pages/publications/105014268058?origin=scopusAI

57. An Integrated Scenario-Based Testing and Explanation Framework for Autonomous Vehicles Goss, Q., Clay Pate, W., Akbas, M.I. Proceedings - 2024 IEEE International Conference on Mobility, Operations, Services and Technologies, MOST 2024, 2024 https://www.scopus.com/pages/publications/85201214304?origin=scopusAI

58. cp3-bench: a tool for benchmarking symbolic regression algorithms demonstrated with cosmology Thing, M.E., Koksbang, S.M. Journal of Cosmology and Astroparticle Physics, 2025 https://www.scopus.com/pages/publications/85215702227?origin=scopusAI

59. Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human Mathew, D.E., Ebem, D.U., Ikegwu, A.C., (...), Dibiaezue, N.F. Neural Processing Letters, 2025 https://www.scopus.com/pages/publications/86000029140?origin=scopusAI

60. Explainable Image Classification: The Journey So Far and the Road Ahead Kamakshi, V., Krishnan, N.C. AI (Switzerland), 2023 https://www.scopus.com/pages/publications/85173106222?origin=scopusAI

61. Explainable AI: Applications, Challenges, Current Solutions and Future Research Directions Butt, T., Iqbal, M. ACM International Conference Proceeding Series, 2025 https://www.scopus.com/pages/publications/105010487424?origin=scopusAI

62. Decoding the Black Box: A Comprehensive Review of Explainable Artificial Intelligence Embarak, O. 2023 9th International Conference on Information Technology Trends, ITT 2023, 2023 https://www.scopus.com/pages/publications/85168553581?origin=scopusAI

63. XAI is in trouble Weber, R.O., Johs, A.J., Goel, P., Silva, J.M. AI Magazine, 2024 https://www.scopus.com/pages/publications/85199996483?origin=scopusAI

64. Measuring Interpretability: A systematic literature review of interpretability measures in artificial intelligence Goel, P., Weber, R. Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS, 2025 https://www.scopus.com/pages/publications/105007944116?origin=scopusAI

65. Revisiting Fidelity in Explainable AI: Unpacking Cognitive Biases and Deceptive Transparency in Model Interpretations Patel, A. 2025 5th Intelligent Cybersecurity Conference, ICSC 2025, 2025 https://www.scopus.com/pages/publications/105016538246?origin=scopusAI

66. Bridging Theory and Application: A Review of Explainable AI with a Case Study in Adaptive Learning Systems Embarak, O. Procedia Computer Science, 2025 https://www.scopus.com/pages/publications/105015141954?origin=scopusAI

67. A Method-Oriented Review of Explainable Artificial Intelligence for Neurological Medical Imaging Peng, C., Li, L., Peng, D. Expert Systems, 2025 https://www.scopus.com/pages/publications/105012724865?

origin=scopusAI

68. A Maturity Model for Practical Explainability in Artificial Intelligence-Based Applications: Integrating Analysis and Evaluation (MM4XAI-AE) Models Muñoz-Ordóñez, J., Cobos, C., Vidal-Rojas, J.C., Herrera, F. International Journal of Intelligent Systems, 2025 https://www.scopus.com/pages/publications/105009270958?origin=scopusAI

69. "how Good Is Your Explanation?": Towards a Standardised Evaluation Approach for Diverse XAI Methods on Multiple Dimensions of Explainability Bhattacharya, A., Verbert, K. UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024 https://www.scopus.com/pages/publications/85198918147?origin=scopusAI

70. Utilizing Federated Learning and SHAP for Predictive Analysis in Smart Grid Security Sudhakara, S.H., Haghnegahdar, L., GhasemiGol, M., Takabi, D. Communications in Computer and Information Science, 2025 https://www.scopus.com/pages/publications/105002577502?origin=scopusAI

71. Federated Learning for Rule-Based Systems: Preliminary Studies Samandari, A., Marchese, M., Paglialonga, A., (...), Mongelli, M. 8th IEEE International Forum on Research and Technologies for Society and Industry Innovation, RTSI 2024 - Proceeding, 2024 https://www.scopus.com/pages/publications/85213818194?origin=scopusAI

72. FedFTL-R: Feature-Interactive Federated Transfer Learning from a Reinforcement Learning Perspective Tong, W., Hu, C., Xie, H. Communications in Computer and Information Science, 2025 https://www.scopus.com/pages/publications/105011364529?origin=scopusAI

73. Blockchain-Enabled Federated Learning Systems with Explainable AI: A Review Potdukhe, N., Gourshettiwar, P., Zade, S., Waghale, A. 2nd International Conference on Machine Learning and Autonomous Systems, ICMLAS 2025 - Proceedings, 2025 https://www.scopus.com/pages/publications/105004801010?origin=scopusAI

74. A statistical package for safe artificial intelligence Babaei, G., Giudici, P. Statistical Methods and Applications, 2025 https://www.scopus.com/pages/publications/105007246865?origin=scopusAI

75. Is a Fairness Metric Score Enough to Assess Discrimination Biases in Machine Learning? Jourdan, F., Pons, R., Asher, N., (...), Risser, L. CEUR Workshop Proceedings, 2023 https://www.scopus.com/pages/publications/85168304725?origin=scopusAI

76. Against generalisation: Data-driven decisions need context to be human-compatible Richardson, S. Business Information Review, 2021 https://www.scopus.com/pages/publications/85120472762?origin=scopusAI

77. AI metrics and policymaking: assumptions and challenges in the shaping of AI Sioumalas-Christodoulou, K., Tympas, A. AI and Society, 2025 https://www.scopus.com/pages/publications/85218814065?origin=scopusAI

78. Is your algorithm dangerous? [Leading edge] Bennett Moses, L. IEEE Technology and Society Magazine, 2018 https://www.scopus.com/pages/publications/85053408185?origin=scopusAI

79. Challenges and Opportunities on the Application of Heuristic Evaluations: A Systematic Literature Review Lecaros, A., Paz, F., Moquillaza, A. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021 https://www.scopus.com/pages/publications/85112200612?origin=scopusAI

80. Can GPT-4o Evaluate Usability Like Human Experts? A Comparative Study on Issue Identification in Heuristic Evaluation Guerino, G., Rodrigues, L., Capeleti, B., (...), Zaina, L. Lecture Notes in Computer Science, 2026 https://www.scopus.com/pages/publications/105017125070?origin=scopusAI

81. Applying heuristic evaluation to human-robot interaction systems Clarkson, E., Arkin, R.C. Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2007, 2007 https://www.scopus.com/pages/publications/37349104941?origin=scopusAI

82. Heuristics of constructing the architecture of an interpreted machine learning model Pylov, P., Dyagileva, A., Protodyakonov, A., Maitak, R. E3S Web of Conferences, 2024
https://www.scopus.com/pages/publications/85196156811?origin=scopusAI

83. A Multidimensional Taxonomy for Recent Trends in Explainable Artificial Intelligence Carvalho, I., Gonçalo Oliveira, H., Silva, C. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) , 2025
https://www.scopus.com/pages/publications/85210237933?origin=scopusAI