Generated by Scopus AI, Mon Nov 10 2025

Inquiries into artificial intelligence interpretability

Machine Learning Interpretability in Artificial Intelligence: Structure, Reasoning, Inference, and Research Gaps

Quick Reference

Key Findings Table

Theme	Summary	Research Gap/Concern	Citations
Intrinsic vs. Post-hoc Interpretability	Distinction between models interpretable by design and those requiring post-hoc explanations.	Need for unified frameworks and rigorous taxonomies; conceptual ambiguity persists.	1 2 3
Accuracy-Interpretability Tradeoff	High accuracy often comes at the cost of interpretability, especially in deep models.	Balancing accuracy and interpretability in high-stakes domains remains unresolved.	4 5 6 7
Evaluation Metrics and Benchmarking	Emergence of standardized platforms (EXACT, XAIB, Co-12 Framework).	Inconsistent protocols and lack of human-centered benchmarks.	8 9 10 11
User-Centric, Context-Aware Explanations	Explanations tailored to user needs, leveraging HCI and cognitive psychology.	Need for interactive, multimodal, and personalized explanation frameworks.	12 13 14
High-Dimensional Data & Feature Dependencies	Challenges in explaining models with many, interdependent features.	Robust methods for grouping, causal modeling, and uncertainty quantification are underdeveloped.	15 16 17 18
Stability & Formal Guarantees	Efforts to ensure explanation robustness (e.g., Multiplicative Smoothing, Lipschitz continuity).	Need for formal guarantees and stability metrics across explanation methods.	19 20 21

Direct Answer

Inquiries into the structure, reasoning, inference, and implications of artificial intelligence using machine learning interpretability methods have focused on differentiating between intrinsically interpretable models and black-box systems requiring post-hoc explanations. Techniques such as saliency maps, attention mechanisms, and concept-based methods have been explored to shed light on deep neural networks, graph neural networks, and transformers. These investigations reveal concerns such as the tradeoff between model accuracy and interpretability, challenges in quantifying explanation fidelity, user-specific explanation needs, and difficulties in handling high-dimensional, dependent features. Significant research gaps remain, including the need for standardized evaluation metrics, improved robustness and stability of explanations, and the integration of user-centered design principles to ensure

that explanations are both technically accurate and actionable for diverse stakeholders. [1] [3] [4] [8] [12] [17] [19] [22]

Study Scope

- **Time Period:** 2018–2024
- **Disciplines:** Machine learning, deep learning, human-computer interaction, cognitive psychology, regulatory science
- **Methods:** Systematic reviews, meta-analyses, empirical benchmarking, user studies, theoretical modeling, computational literature review

Assumptions & Limitations

- Interpretability and explainability are treated as distinct but overlapping concepts.
- Most findings are based on empirical studies and meta-analyses; some theoretical gaps remain.
- Evaluation protocols and datasets are not universally standardized.
- User studies may not generalize across all domains or populations.

Suggested Further Research

- Develop unified, standardized evaluation frameworks for interpretability.
- Advance methods for robust, context-aware, and user-centric explanations.
- Address the accuracy-interpretability tradeoff via allied learning paradigms and neural architecture search.
- Improve handling of high-dimensional, dependent features and uncertainty quantification.
- Integrate regulatory compliance and ethical considerations into interpretability metrics.

1. Introduction

Interpretability in artificial intelligence (AI) is a cornerstone for transparency, trust, and responsible deployment. As AI systems become increasingly complex, understanding their internal structure, reasoning, and inference mechanisms is critical for both theoretical advancement and practical application. Machine learning interpretability methods—ranging from intrinsic model designs to post-hoc explanation techniques—have emerged as essential tools for probing the "black box" of AI, revealing how decisions are made and what features drive predictions. This report synthesizes inquiries into AI structure and reasoning using interpretability methods, identifies key concerns and research gaps, and bridges theoretical foundations with practical insights for future research and deployment [22] [23] [24].

Scope and Motivation

The motivation for investigating interpretability in AI is twofold: (1) to enhance theoretical understanding of model behavior and decision processes, and (2) to ensure practical deployment aligns with ethical, regulatory, and user-centric requirements. Interpretability is especially vital in high-stakes domains such as healthcare, finance, and autonomous systems, where transparency and accountability are paramount [22] [23] [24].

2. Interpretability Methods in AI Structure and Reasoning

Intrinsic vs. Post-hoc Interpretability Frameworks

A foundational distinction in interpretability research is between intrinsic interpretability—models designed to be transparent by construction—and post-hoc explainability, which applies supplementary techniques to elucidate black-box models. Systematic frameworks classify methods along axes such as intrinsic vs. extrinsic, specific vs. agnostic, and local vs. global, aiding in method selection and evaluation [1] [22] [25]. Taxonomies and literature reviews emphasize the need for rigorous classification to resolve conceptual ambiguities and guide practical application.

Interpretability Across Deep Learning Architectures

Interpretability methods vary significantly across deep learning architectures:

- **CNNs:** Explained via activation maps, feature importance, and attention mechanisms. Visual explanations (e.g., heatmaps, class activation maps) are common but challenged by high-dimensional features and complex inference. Probabilistic models (e.g., joint Gaussian mixture models) improve consistency and faithfulness [26] [27] [28].
- **RNNs:** Benefit from graphical representations of hidden state transitions, often compared to finite state automata, enhancing human understanding of sequential data processing [28].
- **Transformers:** Interpreted using attention mechanisms, but individual component analysis can be misleading due to functional equivalence; holistic approaches are needed for faithful interpretation [29] [30] [31].

Quantitative evaluation frameworks are essential to assess the suitability and accuracy of post-hoc methods across architectures, addressing issues like human judgment dependence and data distribution shifts [32] [33].

Advances in Graph Neural Network Interpretability

Recent advances in graph neural network (GNN) interpretability include:

- **Causal Disentanglement:** AMSVGAE-based AMSExplainer provides accurate, compact, and faithful explanations for arbitrary GNN architectures, revealing causal relationships in structural reasoning [34].
- **Meta-path Perturbation:** HGExplainer transforms heterogeneous data for improved trustworthiness in complex graphs [35].
- **Self-Explaining GNNs:** MSE-GNN generates explanations alongside predictions, mimicking human attention in few-shot learning [36].
- **Interactive Tools:** gInterpreter profiles multiple GNN interpretability methods for comparative analysis [37].

- **Global Concept-Based Explanations:** Individual GNN neurons act as detectors of high-level semantic concepts, enabling logically backed global explanations [38].
- **Subgraph Retrieval:** Key subgraph identification methods provide coherent and robust explanations [39].

Hierarchical and Concept-Based Explanations

Interpretability methods increasingly focus on hierarchical reasoning and feature interactions:

- **Hierarchical Fuzzy Systems:** Combine rule-based and structural complexity measures to capture multi-layered reasoning [40].
- **Feature Interaction Attribution:** Reveal hierarchical representations and grammatical rules in language models [41].
- **Concept-Supported XAI:** Emphasize higher-level, human-understandable concepts, aligning explanations with human cognition [18].
- **Prototype-Based Explainers:** HIPE leverages hierarchical relationships for multi-level explanations [42].
- **Automatic Concept Extraction:** Methods like ACE systematically summarize feature interactions and hierarchical reasoning [43].

Synthesis: Interpretability methods have provided structured insights into AI reasoning, enabling differentiation between model types, revealing hierarchical and causal relationships, and supporting practical deployment in sensitive domains. However, challenges remain in adapting methods to complex architectures and ensuring explanation fidelity [1] [26] [34] [41].

3. Theoretical and Practical Concerns in AI Interpretability

Accuracy-Interpretability Tradeoff

A central concern is the tradeoff between model accuracy and interpretability. High-performance models (e.g., deep neural networks, transformers) are often opaque, while interpretable models may sacrifice predictive accuracy. This tradeoff is particularly acute in high-stakes domains, where transparency and accountability are essential [4] [7] [44] [45]. Ethical and regulatory frameworks are evolving to address this tension, recommending stronger obligations for explainability and transparency.

User-Centric Explanation Design

Effective interpretability requires explanations tailored to user needs:

- **Causal, Contrastive, Contextual Explanations:** Interactive selection and user feedback improve trust and comprehension [12] [46] [47].
- **Cognitive and HCI Principles:** Visual hierarchy, consistency, and adaptive interfaces reduce cognitive load and biases, enhancing user experience [48] [49].

- **Multimodal Modalities:** Combining textual, visual, and audio explanations increases satisfaction and recall, especially for non-expert users [47] [50].

Computational Efficiency and Scalability

Popular interpretability tools (e.g., SHAP, LIME) face computational challenges:

- **SHAP:** Provides robust global explanations but incurs high computational overhead, limiting scalability in real-time systems [51].
- **LIME:** Offers faster, local explanations but can be unstable due to instance perturbations [51].
- **Hybrid Approaches:** Combining SHAP and LIME balances efficiency and interpretability for real-time deployment [52].

Conceptual Ambiguities and Evaluation Challenges

Ambiguities between interpretability and explainability complicate method selection and evaluation:

- **Conceptual Distinction:** Interpretability relates to inherent model transparency; explainability concerns human understanding [1] [3].
- **Evaluation Metrics:** Lack of consensus on objective measurement impedes systematic progress; multi-faceted conceptualization is needed [10] [25].

Synthesis: Theoretical and practical concerns center on balancing accuracy with interpretability, designing user-centric explanations, ensuring computational efficiency, and resolving conceptual ambiguities. Addressing these concerns is critical for trustworthy and actionable AI [3] [7] [12] [51].

4. Limitations and Challenges in Interpretability Approaches

Biases in Feature Importance and Surrogate Models

Feature importance and surrogate model-based methods have inherent biases:

- **Encoding Effects:** Bias levels vary with feature encoding (numerical vs. categorical) [53].
- **Cutoff Thresholds:** Lack of consensus on feature selection thresholds limits interpretability [54].
- **Stability Guarantees:** Smoothing techniques (e.g., Multiplicative Smoothing) improve stability but may not fully reflect true decision processes [19].
- **Surrogate Fidelity:** Surrogate models may lose accuracy and interpretability in high-dimensional systems; careful validation is required [55] [56].

Handling Dependent Features and Uncertainty

Interpretability methods often assume feature independence, leading to unrealistic explanations in scientific applications:

- **Group Perturbation:** Modified SHAP/LIME for spectral zones better captures dependencies [17] [18].
- **Uncertainty Quantification:** Critical for reliable decision-making; integrated visualization interfaces aid domain experts [57] [58].

Formalized Datasets and Benchmarking

Lack of standardized, interpretable datasets hampers validation and comparison:

- **GWAP Platforms:** Collect human evaluations for benchmarking (e.g., Eye into AI) [59].
- **Co-12 Framework:** Comprehensive property assessment for explanation quality [10].
- **Domain-Specific Surveys:** Highlight need for formalized datasets in areas like plant health monitoring [60].

Conceptual and Methodological Fragmentation

Inconsistent classification between interpretability and explainability hinders unified guidelines:

- **Conceptual Ambiguity:** Miscommunication and fragmented approaches result from conflating concepts [1] [2].
- **Regulatory Impact:** Disparate interpretations affect policy and accountability [61] [62].

Synthesis: Limitations in current interpretability techniques include biases, challenges with dependent features, lack of standardized datasets, and conceptual fragmentation. Addressing these issues is essential for advancing reliable and actionable interpretability [2] [17] [53] [59].

5. Improving Stability, Faithfulness, and Fairness in Interpretability

Formal Guarantees for Feature Attribution Methods

Efforts to improve stability and faithfulness include:

- **SHAP-Guided Regularization:** Entropy-based penalties enhance generalization and explanation reliability [63].
- **Multiplicative Smoothing (MuS):** Enforces Lipschitz continuity for formal stability guarantees, applicable to SHAP and LIME [19].
- **Distribution-Aware Methods:** SHAP-KL and FastSHAP-KL maintain label distribution fidelity [64].
- **Stability Metrics:** Quantitative evaluation reveals SHAP provides more stable explanations than LIME [20].

Mitigating Biases in Data Representation and Encoding

Feature encoding and data representation biases affect interpretability and fairness:

- **Encoding-Induced Bias:** One-hot and target encoding influence model performance and fairness; regularization mitigates some unfairness [65].
- **Data Representation Bias:** Label, selection, and underrepresentation biases propagate discrimination; synthetic data augmentation and expert involvement improve fairness [66] [67].

Limitations and Improvements in Surrogate Modeling

Surrogate models face trade-offs between efficiency and fidelity:

- **Dimensionality Reduction:** Kernel principal component analysis enables high-dimensional surrogate modeling [68].
- **Hybrid Models:** Combine different fidelity data and adaptive sampling for robustness [69].
- **Uncertainty Quantification:** Bayesian frameworks and ensembling improve reliability [70].
- **Joint Training:** Multi-objective optimization enhances surrogate fidelity [71].

Aggregation and Smoothing for Explanation Stability

Aggregation and smoothing techniques improve explanation stability:

- **Global Cluster-Based Strategies:** Reduce disagreement and instability in explanations [72].
- **Distribution Compression:** "Compress then explain" (CTE) reduces approximation error and computational cost [73].
- **Weighted Importance Score:** Aggregates feature rankings for reproducibility [74].
- **Federated Learning Aggregation:** Improves explanation fidelity under non-IID data [75].

Synthesis: Formal guarantees, bias mitigation, surrogate model improvements, and aggregation techniques collectively enhance the stability, faithfulness, and fairness of interpretability methods, supporting more reliable and equitable AI systems [63] [65] [68] [72].

6. Future Research Directions in AI Interpretability

Standardized Evaluation Metrics and Frameworks

Proposed metrics and platforms include:

- **XAIB:** Universal, extensible platform with Co-12 Framework for comprehensive evaluation [8].
- **EXACT:** Benchmarking with ground truth explanations and quantitative metrics [76].
- **Quantitative Indices:** Consistency, user comprehension, causality, effectiveness, stability [77].

- **N-IOU and N-Recall:** Measure explanation accuracy and recall [78].

Optimizing Accuracy-Interpretability Tradeoff

Allied learning paradigms and neural architecture search offer solutions:

- **Explanation-Guided Training:** Focuses model training on salient input regions [79].
- **Prototype-Based Architectures:** Provide human-understandable explanations with strong accuracy [80].
- **Hybrid Models:** Combine interpretable components with deep learning [81].
- **Intrinsic Interpretability:** Designing models for real-time, accurate, and actionable explanations [82].

Multimodal and User-Tailored Explainability Frameworks

Frameworks adapt explanations to diverse user profiles:

- **Social Explainable AI (sXAI):** Interactive, context-aware explanations using knowledge graphs [14].
- **Dual XAI:** Adapts content and format to user expertise [83].
- **TELL-ME:** Personalized explanations for large language models [84].
- **ConEX:** Context-sensitive explanations for recommender systems [85].

Computational Literature Review and Gap Identification

AI-driven literature review methods aid gap identification:

- **AI-Specific PLMs:** Dynamic mapping of research landscapes [86].
- **Hybrid Frameworks:** Integrate computational power with human expertise [87].
- **Interactive Topic Evolution Maps:** Track and anticipate shifts in interpretability research [86].
- **Bibliometric Analyses:** Reveal evolving hotspots and collaboration networks [88].

Synthesis: Future research should focus on standardizing evaluation, optimizing the accuracy-interpretability tradeoff, developing multimodal and user-tailored frameworks, and leveraging computational literature review methods to identify and address evolving research gaps [7] [8] [14] [86].

7. Benchmarking and Regulatory Compliance in Interpretability

Quantitative Metrics and Benchmarking Protocols

Effective metrics and protocols include:

- **Feature Synergy:** Captures strength of feature interactions [9].

- **Model-Agnostic Complexity Measures:** Quantify model complexity and reliability [89].
- **BEExAI:** Large-scale comparison of post-hoc methods [90].
- **RemOve And Retrain (ROAR):** Evaluates explanation quality in image classification [33].

Explainability Metrics for Regulatory Compliance

Designing metrics for compliance:

- **EU AI Act:** Requires context-dependent, stakeholder-tailored explanations [91].
- **Explanation Groves:** Balance complexity and degree of explanation [92].
- **Human-Centered Design:** Ensures explanations are understandable and actionable [93].
- **Taxonomies:** Integrate legal, end-user, and engineer perspectives [94].

Comparative Analysis of Benchmarking Platforms

Platforms differ in evaluation ontologies:

- **EXACT:** Empirical benchmarking with ground truth and quantitative metrics [76].
- **LATEC:** Evaluates multiple metrics and model variations [95].
- **Gaps:** Lack of universally accepted protocols, inconsistent metrics, insufficient human-centered frameworks [96] [97].

Synthesis: Quantitative metrics, benchmarking protocols, and regulatory compliance frameworks are essential for evaluating interpretability methods and ensuring actionable, trustworthy AI. Standardization and human-centered approaches remain key challenges [9] [76] [91].

8. Conclusion

Summary of Insights and Research Gaps

The landscape of AI interpretability is multifaceted, interweaving theoretical, methodological, and practical concerns. Key findings include:

- **Differentiation between intrinsic and post-hoc interpretability** is foundational for method selection and evaluation.
- **Accuracy-interpretability tradeoff** remains a central challenge, especially in high-stakes domains.
- **Standardized evaluation metrics and benchmarking platforms** are emerging but lack universal adoption.
- **User-centric, context-aware explanation design** is critical for trust and actionable insights.

- **Handling high-dimensional, dependent features and uncertainty** requires robust, adaptive methods.
- **Formal guarantees, stability, and fairness** are essential for reliable interpretability.
- **Regulatory compliance and ethical considerations** must be integrated into interpretability metrics.

Persistent research gaps include the need for unified frameworks, improved robustness and stability, context-sensitive and user-tailored explanations, and comprehensive evaluation protocols. Interdisciplinary collaboration and continued innovation are vital for advancing interpretability in AI, ensuring systems are not only powerful but also transparent, trustworthy, and aligned with societal values [24] [98] [99].

End of Report

References

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

2. Interpretability versus Explainability: Classification for Understanding Deep Learning Systems and Models Namatēvs, I., Sudars, K., Dobrājs, A. Computer Assisted Methods in Engineering and Science, 2022 https://www.scopus.com/pages/publications/85144533773?origin=scopusAI

3. 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

4. Interpretation of Artificial Intelligence Algorithms in the Prediction of Sepsis Murugesan, I., Murugesan, K., Balasubramanian, L., Arumugam, M. Computing in Cardiology, 2019 https://www.scopus.com/pages/publications/85081116131?origin=scopusAI

5. 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

6. Boosting Human Competences With Interpretable and Explainable Artificial Intelligence Herzog, S.M., Franklin, M. Decision, 2024 https://www.scopus.com/pages/publications/85208276865?origin=scopusAI

7. 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

8. 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

9. An evolutionary approach to interpretable learning Robertson, J., Hu, T. GECCO 2021 Companion - Proceedings of the 2021 Genetic and Evolutionary Computation Conference Companion, 2021 https://www.scopus.com/pages/publications/85111008803?origin=scopusAI

10. 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

11. How do ML practitioners perceive explainability? an interview study of practices and challenges Habiba, U.-E., Habib, M.K., Bogner, J., (...), Wagner, S. Empirical Software Engineering, 2025 https://www.scopus.com/pages/publications/85208552436?origin=scopusAI

12. A position on establishing effective explanations from human-centred counterfactuals for automated financial decisions Blandin, A., Roach, M., Doneddu, D., (...), Sullivan, D. Proceedings of the AISB Convention 2023, 2023 https://www.scopus.com/pages/publications/85176318049?origin=scopusAI

13. iSee: Advancing Multi-Shot Explainable AI Using Case-Based Recommendations Wijekoon, A., Wiratunga, N., Corsar, D., (...), Liret, A. Frontiers in Artificial Intelligence and Applications, 2024 https://www.scopus.com/pages/publications/85216660296?origin=scopusAI

14. Social Explainable AI: What Is It and How to Make It Happen with CIU? Främling, K. Lecture Notes in Computer Science, 2026 https://www.scopus.com/pages/publications/105020008823?origin=scopusAI

15. GEASS: NEURAL CAUSAL FEATURE SELECTION FOR HIGH-DIMENSIONAL BIOLOGICAL DATA Dong, M., Kluger, Y. 11th International Conference on Learning Representations, ICLR 2023, 2023 https://www.scopus.com/pages/publications/85199918552?origin=scopusAI

16. CLASS SPECIFIC INTERPRETABILITY IN CNN USING CAUSAL ANALYSIS Yadu, A., Suhas, P.K., Sinha, N. Proceedings - International Conference on Image Processing, ICIP, 2021 https://www.scopus.com/pages/publications/85125585103?origin=scopusAI

17. Spectral Zones-Based SHAP/LIME: Enhancing Interpretability in Spectral Deep Learning Models Through Grouped Feature Analysis Contreras, J., Winterfeld, A., Popp, J., Bocklitz, T. Analytical Chemistry, 2024 https://www.scopus.com/pages/publications/85205723334?origin=scopusAI

18. Navigating the landscape of concept-supported XAI: Challenges, innovations, and future directions Shams Khoozani, Z., Sabri, A.Q.M., Seng, W.C., (...), Eg, K.Y. Multimedia Tools and Applications, 2024 https://www.scopus.com/pages/publications/85182636559?origin=scopusAI

19. Stability Guarantees for Feature Attributions with Multiplicative Smoothing Xue, A., Alur, R., Wong, E. Advances in Neural Information Processing Systems, 2023 https://www.scopus.com/pages/publications/85191157917?origin=scopusAI

20. Towards Reliable Explainable AI: A Novel Stability Metric for Trustworthy Interpretations Butt, T.A., Iqbal, M. Lecture Notes in Networks and Systems, 2025 https://www.scopus.com/pages/publications/105019530467?origin=scopusAI

21. Locally Invariant Explanations: Towards Stable and Unidirectional Explanations through Local Invariant Learning Dhurandhar, A., Ramamurthy, K.N., Ahuja, K., Arya, V. Advances in Neural Information Processing Systems, 2023 https://www.scopus.com/pages/publications/85186317881?origin=scopusAI

22. Interpretable and explainable machine learning methods for predictive process monitoring: a systematic literature review Mehdiyev, N., Majlatow, M., Fettke, P. Artificial Intelligence Review, 2025 https://www.scopus.com/pages/publications/105018851066?origin=scopusAI

23. Enhancing Project Security: Unveiling Trust, Interpretability and Explainability in the Age of AI Alshar'e, M., Abualkishik, A., Abuhmaidan, K., Kayed, A. 5G-Enabled Technology for Smart City and Urbanization System, 2024 https://www.scopus.com/pages/publications/85214156082?origin=scopusAI

24. Problems of Interpretability and Transparency of Decisions Made by AI Orobinskaya, V.N., Mishina, T.N., Mazurenko, A.P., Mishin, V.V. Proceedings - 2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency, SUMMA 2024, 2024 https://www.scopus.com/pages/publications/105014174882?origin=scopusAI

25. 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

26. Neural Network Interpretability: Methods for Understanding and Visualizing Deep Learning Models Sudhakar, A.V.V., Sharma, M.K., Mohan, C.R., (...), Geetha, B.T. Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024, 2024 https://www.scopus.com/pages/publications/85217388743?origin=scopusAI

27. Joint Gaussian mixture model for versatile deep visual model explanation Xie, Z., He, T., Tian, S., (...), Chen, D. Knowledge-Based Systems, 2023 https://www.scopus.com/pages/publications/85172394399?origin=scopusAI

28. Post hoc explanations for RNNs using state transition representations for time series data Gupta, G. CEUR Workshop Proceedings, 2023 https://www.scopus.com/pages/publications/85178648723?origin=scopusAI

29. Enhancing Robustness and Interpretability in Transformer Networks through Fuzzification Cui, J., Yang, J., Gao, T. 2024 7th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2024, 2024 https://www.scopus.com/pages/publications/105018214610?origin=scopusAI

30. Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars Wen, K., Liu, B., Li, Y., Risteski, A. Advances in Neural Information Processing Systems, 2023 https://www.scopus.com/pages/publications/85178048906?origin=scopusAI

31. Do Vision Transformers See Like Convolutional Neural Networks? Raghu, M., Unterthiner, T., Kornblith, S., (...), Dosovitskiy, A. Advances in Neural Information Processing Systems, 2021 https://www.scopus.com/pages/publications/85131797911?origin=scopusAI

32. Evaluation of post-hoc interpretability methods in time-series classification Turbé, H., Bjelogrlic, M., Lovis, C., Mengaldo, G. Nature Machine Intelligence, 2023 https://www.scopus.com/pages/publications/85149863240?origin=scopusAI

33. Evaluation of Post-hoc Interpretability Methods in Breast Cancer Histopathological Image Classification Waqas, M., Maul, T., Ahmed, A., Liao, I.Y. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2024 https://www.scopus.com/pages/publications/85195140316?origin=scopusAI

34. AMSVGAE-Based Causal Inference for Interpretable Graph Neural Networks Wu, J., Lin, L., Huang, Y., (...), Jia, X. Communications in Computer and Information Science, 2026 https://www.scopus.com/pages/publications/105019505945?origin=scopusAI

35. HGExplainer: Toward Interpretable Heterogeneous Graph Neural Networks via Meta-path Perturbation Wang, Y., Zhang, Z., Yin, J., (...), Dong, B. Journal of Signal Processing Systems, 2025 https://www.scopus.com/pages/publications/105012407791?origin=scopusAI

36. Towards Few-Shot Self-explaining Graph Neural Networks Peng, J., Liu, Q., Yue, L., (...), Sha, Y. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2024 https://www.scopus.com/pages/publications/85203843937?origin=scopusAI

37. A Demonstration of Interpretability Methods for Graph Neural Networks Mobaraki, E.B., Khan, A. ACM International Conference Proceeding Series, 2023 https://www.scopus.com/pages/publications/85163724299?origin=scopusAI

38. Global Concept-Based Interpretability for Graph Neural Networks via Neuron Analysis Xuanyuan, H., Barbiero, P., Georgiev, D., (...), Liò, P. Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, 2023 https://www.scopus.com/pages/publications/85168255076?origin=scopusAI

39. A graph neural network explainability strategy driven by key subgraph connectivity Dai, L.N., Xu, D.H., Gao, Y.F. Journal of Biomedical Informatics, 2025 https://www.scopus.com/pages/publications/105000938317?origin=scopusAI

40. An improved complexity measure in hierarchical fuzzy systems Razak, T.R., Garibaldi, J.M., Wagner, C. IEEE International Conference on Fuzzy Systems, 2020 https://www.scopus.com/pages/publications/85090498279?origin=scopusAI

41. Feature Interactions Reveal Linguistic Structure in Language Models Jumelet, J., Zuidema, W. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2023 https://www.scopus.com/pages/publications/85174872765?origin=scopusAI

42. Hierarchical Explanations for Video Action Recognition Gulshad, S., Long, T., Van Noord, N. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2023 https://www.scopus.com/pages/publications/85170825665?origin=scopusAI

43. Towards automatic concept-based explanations Ghorbani, A., Wexler, J., Zou, J., Kim, B. Advances in Neural Information Processing Systems, 2019 https://www.scopus.com/pages/publications/85087651641?origin=scopusAI

44. “Just” accuracy? Procedural fairness demands explainability in AI-based medical resource allocations Rueda, J., Rodríguez, J.D., Jounou, I.P., (...), Rodríguez-Arias, D. AI and Society, 2024 https://www.scopus.com/pages/publications/85144538125?origin=scopusAI

45. The false hope of current approaches to explainable artificial intelligence in health care Ghassemi, M., Oakden-Rayner, L., Beam, A.L. The Lancet Digital Health, 2021 https://www.scopus.com/pages/publications/85121190788?origin=scopusAI

46. The role of user feedback in enhancing understanding and trust in counterfactual explanations for explainable AI Suffian, M., Kuhl, U., Bogliolo, A., Alonso-Moral, J.M. International Journal of Human Computer Studies, 2025 https://www.scopus.com/pages/publications/105000535783?origin=scopusAI

47. Dynamic attention-based explainable recommendation with textual and visual fusion Liu, P., Zhang, L., Gulla, J.A. Information Processing and Management, 2020 https://www.scopus.com/pages/publications/85070795662?origin=scopusAI

48. Perception-Centric Explainable AI: Bridging Cognitive Theories and HCI Design for Enhanced User Experience Alhasan, S., Alnanih, R. Procedia Computer Science, 2025 https://www.scopus.com/pages/publications/105005183111?origin=scopusAI

49. Design and development of applications using human-computer interaction Pandey, A., Panday, S.P., Joshi, B. Innovations in Artificial Intelligence and Human-Computer Interaction in the Digital Era, 2023 https://www.scopus.com/pages/publications/85175387836?origin=scopusAI

50. Adapting Online Patient Decision Aids: Effects of Modality and Narration Style on Patients’ Satisfaction, Information Recall and Informed Decision Making De Looper, M., Damman, O., Smets, E., (...), Van Weert, J. Journal of Health Communication, 2020 https://www.scopus.com/pages/publications/85097031452?origin=scopusAI

51. A Comparative Study of LIME and SHAP for Enhancing Trustworthiness and Efficiency in Explainable AI Systems Roshinta, T.A., Gábor, S. 2024 IEEE International Conference on Computing, ICOCO 2024, 2024 https://www.scopus.com/pages/publications/105002049557?origin=scopusAI

52. A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration Tahir, H.A., Alayed, W., Hassan, W.U., Haider, A. Sensors, 2024 https://www.scopus.com/pages/publications/85208588366?origin=scopusAI

53. Measuring Implicit Bias Using SHAP Feature Importance and Fuzzy Cognitive Maps Grau, I., Nápoles, G., Hoitsma, F., (...), Vanhoof, K. Lecture Notes in Networks and Systems, 2024 https://www.scopus.com/pages/publications/85182512259?origin=scopusAI

54. Threshold benchmarking for feature ranking techniques Malhotra, R., Sharma, A. Bulletin of Electrical Engineering and Informatics, 2021 https://www.scopus.com/pages/publications/85102989245?origin=scopusAI

55. Efficient Surrogate Modeling Based on Improved Vision Transformer Neural Network for History Matching Zhang, D., Li, H. SPE Journal, 2023 https://www.scopus.com/pages/publications/85182453325?origin=scopusAI

56. Surrogate-Assisted Combinatorial Optimization of EV Fast Charging Stations Lin, J., Gebbran, D., Dragicevic, T. IEEE Transactions on Transportation Electrification, 2024 https://www.scopus.com/pages/publications/85153352303?origin=scopusAI

57. Uncertainty quantification for deep learning in geoscience applications Mosser, L., Purves, S., Zabihi Naeini, E. EAGE/AAPG Digital Subsurface for Asia Pacific Conference 2020, 2020 https://www.scopus.com/pages/publications/85111052614?origin=scopusAI

58. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs Suresh, H., Lewis, K.M., Guttag, J., Satyanarayan, A. International Conference on Intelligent User Interfaces, Proceedings IUI, 2022 https://www.scopus.com/pages/publications/85127820200?origin=scopusAI

59. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose Morrison, K., Jain, M., Hammer, J., Perer, A. Proceedings of the ACM on Human-Computer Interaction, 2023 https://www.scopus.com/pages/publications/85174490262?origin=scopusAI

60. A systematic survey on explainable artificial intelligence (XAI) for plant health monitoring: challenges and opportunities Kaler, B., Kaur, A. Applied Intelligence, 2025 https://www.scopus.com/pages/publications/105011754476?origin=scopusAI

61. A Scoping Review of Transparency and Explainability in AI Ethics Guidelines Hooper, K., Lunn, S. Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS, 2024 https://www.scopus.com/pages/publications/85200436841?origin=scopusAI

62. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK Nannini, L., Balayn, A., Smith, A.L. ACM International Conference Proceeding Series, 2023 https://www.scopus.com/pages/publications/85163698693?origin=scopusAI

63. SHAP-Guided Regularization in Machine Learning Models Saadallah, A. CEUR Workshop Proceedings, 2025 https://www.scopus.com/pages/publications/105015647555?origin=scopusAI

64. Don't be fooled: label leakage in explanation methods and the importance of their quantitative evaluation Jethani, N., Saporta, A., Ranganath, R. Proceedings of Machine Learning Research, 2023 https://www.scopus.com/pages/publications/85165182308?origin=scopusAI

65. Fairness Implications of Encoding Protected Categorical Attributes Mougan, C., Alvarez, J.M., Ruggieri, S., Staab, S. AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023 https://www.scopus.com/pages/publications/85173630317?origin=scopusAI

66. Bias-Tolerant Fair Classification Zhang, Y., Zhou, F., Li, Z., (...), Chen, F. Proceedings of Machine Learning Research, 2021 https://www.scopus.com/pages/publications/85159962393?origin=scopusAI

67. Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness? Duong, M.K., Conrad, S. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2023 https://www.scopus.com/pages/publications/85172339699?origin=scopusAI

68. Extending classical surrogate modeling to high dimensions through supervised dimensionality reduction: A data-driven approach Lataniotis, C., Marelli, S., Sudret, B. International Journal for Uncertainty Quantification, 2020 https://www.scopus.com/pages/publications/85082507147?origin=scopusAI

69. Effects of Surrogate Hybridization and Adaptive Sampling for Simulation-Based Optimization Ravutla, S., Bai, A., Realff, M.J., Boukouvala, F. Industrial and Engineering Chemistry Research, 2025 https://www.scopus.com/pages/publications/105002783793?origin=scopusAI

70. Uncertainty in the era of machine learning for atomistic modeling Grasselli, F., Chong, S., Kapil, V., (...), Rossi, K. Digital Discovery, 2025 https://www.scopus.com/pages/publications/105018114443?origin=scopusAI

71. Joint Explainability-Performance Optimization with Surrogate Models for AI-Driven Edge Services Charalampakos, F., Tsouparopoulos, T., Koutsopoulos, I. 2025 IEEE International Conference on Machine Learning for Communication and Networking, ICMLCN 2025, 2025 https://www.scopus.com/pages/publications/105016778003?origin=scopusAI

72. Clarity in complexity: how aggregating explanations resolves the disagreement problem Mitruț, O., Moise, G., Moldoveanu, A., (...), Petrescu, L. Artificial Intelligence Review, 2024 https://www.scopus.com/pages/publications/85207069493?origin=scopusAI

73. EFFICIENT AND ACCURATE EXPLANATION ESTIMATION WITH DISTRIBUTION COMPRESSION Baniecki, H., Casalicchio, G., Bischl, B., Biecek, P. 13th International Conference on Learning Representations, ICLR 2025, 2025 https://www.scopus.com/pages/publications/105010280006?origin=scopusAI

74. Synthesizing Explainability Across Multiple ML Models for Structured Data Veledar, E., Zhou, L., Veledar, O., (...), Rundek, T. Algorithms, 2025 https://www.scopus.com/pages/publications/105009270527?origin=scopusAI

75. SPATL-XLC: An Explainability-Driven Framework for Efficient and Robust Federated Learning Under Non-IID Data Seifu, S.H., Assefa, B.G. IEEE Access, 2025 https://www.scopus.com/pages/publications/105011065361?origin=scopusAI

76. 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

77. Quantitative evaluation method for interpretability of XAI based on surrogate model Li, Y., Wang, C.-L., Zuo, X.-Q., (...), Zhang, X.-J. Kongzhi yu Juece/Control and Decision, 2024 https://www.scopus.com/pages/publications/85184061146?origin=scopusAI

78. Research on Quantitative Evaluation Method of Interpretability Based on Deep Learning Model An, S., Wu, Y., Bai, Y. 2024 4th International Symposium on Artificial Intelligence and Intelligent Manufacturing, AIIM 2024, 2024 https://www.scopus.com/pages/publications/105002237326?origin=scopusAI

79. ET: EXPLAIN TO TRAIN: LEVERAGING EXPLANATIONS TO ENHANCE THE TRAINING OF A MULTIMODAL TRANSFORMER Ayyar, M.P., Benois-Pineau, J., Zemmari, A. Proceedings - International Conference on Image Processing, ICIP, 2024 https://www.scopus.com/pages/publications/85216875917?origin=scopusAI

80. ENHANCED PROTOTYPICAL PART NETWORK (EPPNET) FOR EXPLAINABLE IMAGE CLASSIFICATION VIA PROTOTYPES Atote, B., Sanchez, V. Proceedings - International Conference on Image Processing, ICIP, 2024 https://www.scopus.com/pages/publications/85216890899?origin=scopusAI

81. An End-to-End Trainable Deep Convolutional Neuro-Fuzzy Classifier Yeganejou, M., Kluzinski, R., Dick, S., Miller, J. IEEE International Conference on Fuzzy Systems, 2022 https://www.scopus.com/pages/publications/85138799603?origin=scopusAI

82. Viewpoint: The Future of Human-Centric Explainable Artificial Intelligence is not Post-Hoc Explanations Swamy, V., Frej, J., Käser, T. Journal of Artificial Intelligence Research, 2025 https://www.scopus.com/pages/publications/105015296225?origin=scopusAI

83. Beyond Model Trust: Dual XAI for Adaptive and User-Centric Explainability De Bonis, M.L.N. CEUR Workshop Proceedings, 2025 https://www.scopus.com/pages/publications/105015753119?origin=scopusAI

84. TELL-ME: Toward Personalized Explanations of Large Language Models Jeck, J., Leiser, F., Hüsges, A., Sunyaev, A. Conference on Human Factors in Computing Systems - Proceedings , 2025 https://www.scopus.com/pages/publications/105005732043?origin=scopusAI

85. ConEX: A Context-Aware Framework for Enhancing Explanation Systems Khaled, Y., Ehab, N. International Conference on Agents and Artificial Intelligence, 2024 https://www.scopus.com/pages/publications/85190785416?origin=scopusAI

86. Harnessing language models for computational literature review of emerging AI topics Chung, J., Jeong, B., Park, Y.-J., (...), Choi, J. Information Processing and Management, 2025 https://www.scopus.com/pages/publications/105007517225?origin=scopusAI

87. A hybrid framework for creating artificial intelligence-augmented systematic literature reviews Malik, F.S., Terzidis, O. Management Review Quarterly, 2025 https://www.scopus.com/pages/publications/105003873305?origin=scopusAI

88. Bibliometric Analysis of Large Language Model Artificial Intelligence Based on Knowledge Graph Weng, X., Wang, Y., Weng, S., (...), Weng, L. Proceeding - 2024 IEEE 9th International Conference on Data Science in Cyberspace, DSC 2024, 2024 https://www.scopus.com/pages/publications/85218444659?origin=scopusAI

89. Quantifying model complexity via functional decomposition for better post-hoc interpretability Molnar, C., Casalicchio, G., Bischl, B. Communications in Computer and Information Science, 2020 https://www.scopus.com/pages/publications/85083706831?origin=scopusAI

90. BEExAI: Benchmark to Evaluate Explainable AI Sithakoul, S., Meftah, S., Feutry, C. Communications in Computer and Information Science, 2024 https://www.scopus.com/pages/publications/85200775981?origin=scopusAI

91. Algorithmic Knowability: A Unified Approach to Explanations in the AI Act Sapienza, S., Palmirani, M. Communications in Computer and Information Science, 2026 https://www.scopus.com/pages/publications/105020239495?origin=scopusAI

92. Explanation Groves – Controlling the Trade-off between the Degree of Explanation vs. its Complexity Szepannek, G. CEUR Workshop Proceedings, 2025 https://www.scopus.com/pages/publications/105015598725?origin=scopusAI

93. The European commitment to human-centered technology: The integral role of HCI in the EU AI Act's success Calero Valdez, A., Heine, M., Franke, T., (...), Schrills, T. i-com, 2024 https://www.scopus.com/pages/publications/85198930848?origin=scopusAI

94. Towards Explainability as a Functional Requirement: A Vision to Integrate the Legal, End-User, and ML Engineer Perspectives Habiba, U.-E.-., Bogner, J., Wagner, S. Proceedings - 2024 IEEE/ACM International Workshop on Responsible AI Engineering, RAIE 2024, 2024 https://www.scopus.com/pages/publications/85201227070?origin=scopusAI

95. Navigating the Maze of Explainable AI: A Systematic Approach to Evaluating Methods and Metrics Klein, L., Lüth, C., Schlegel, U., (...), Jäger, P. Advances in Neural Information Processing Systems, 2024 https://www.scopus.com/pages/publications/105000547444?origin=scopusAI

96. Not all explanations are created equal: investigating the pitfalls of current XAI evaluation Shymanski, J., Brue, J., Sen, S. Bi-directionality in Human-AI Collaborative Systems, 2025 https://www.scopus.com/pages/publications/105019733308?origin=scopusAI

97. Human-centered evaluation of explainable AI applications: a systematic review Kim, J., Maathuis, H., Sent, D. Frontiers in Artificial Intelligence, 2024 https://www.scopus.com/pages/publications/85208614239?origin=scopusAI

98. Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research Brdnik, S., Šumak, B. 2024 47th ICT and Electronics Convention, MIPRO 2024 - Proceedings, 2024 https://www.scopus.com/pages/publications/85198223172?origin=scopusAI

99. Tertiary Review on Explainable Artificial Intelligence: Where Do We Stand? van Mourik, F., Jutte, A., Berendse, S.E., (...), Ahmed, F. Machine Learning and Knowledge Extraction, 2024 https://www.scopus.com/pages/publications/85205242087?origin=scopusAI