

Explainable AI Models for Transparent and Responsible Automated Decision-Making

Kary Främling

Professor, Umeå University, Sweden

Abstract

The rapid advancement of Artificial Intelligence (AI) has significantly transformed automated decision-making across healthcare, finance, education, public administration, and industrial management. AI-driven systems are increasingly being used to analyse complex datasets, identify patterns, and generate predictive outcomes that support strategic and operational decisions. However, the growing dependence on automated decision systems has created concerns regarding transparency, accountability, fairness, and user trust. Many advanced AI models, particularly deep learning architectures, function as complex black-box systems where decision processes are difficult for human users to understand. This limitation restricts their acceptance in high-impact domains where explanations and responsible decision-making are essential.

Explainable Artificial Intelligence (XAI) has emerged as an important research direction that aims to improve the interpretability and transparency of AI models while maintaining predictive performance. This study examines explainable AI models designed to support transparent and responsible automated decision-making by integrating interpretability techniques, ethical principles, and human-centered evaluation approaches. A simulation-based analytical framework is developed to compare conventional AI models with explainable AI approaches using performance indicators including prediction accuracy, explanation quality, transparency level, and user trust.

The findings indicate that explainable AI models provide improved transparency and stronger decision reliability by enabling users to understand the reasoning behind automated recommendations. The study highlights that responsible AI adoption requires a balanced approach where technical performance is combined with accountability, fairness, and meaningful human oversight. The proposed framework contributes to ongoing discussions on trustworthy artificial intelligence by presenting explainability as a fundamental requirement for future automated decision systems.

Keywords: Explainable Artificial Intelligence, Transparent AI, Responsible AI, Automated Decision-Making, AI Governance, Machine Learning Interpretability, Trustworthy Artificial Intelligence.

1. Introduction

Artificial Intelligence has become one of the most influential technologies shaping modern decision-making environments by enabling automated systems to analyse large volumes of data, recognize complex patterns, and generate predictive insights. Organizations across different sectors increasingly rely on AI-based solutions to improve operational efficiency, optimize resource allocation, and support

strategic planning. Machine learning and deep learning models have demonstrated remarkable capabilities in areas such as medical diagnosis, financial forecasting, fraud detection, recommendation systems, and intelligent automation. However, the increasing complexity of these models has created significant concerns regarding the ability of users to understand how automated decisions are generated.

The limited interpretability of advanced AI models has introduced challenges related to transparency, accountability, and trust. Many high-performing models operate as black-box systems where internal decision processes remain difficult to explain even when their predictions are accurate. In sensitive applications, automated decisions can directly influence individuals, organizations, and communities, making it essential to understand the reasoning behind AI-generated outcomes. The absence of clear explanations may reduce user confidence, create ethical concerns, and increase difficulties in identifying potential biases or errors within AI systems.

Explainable Artificial Intelligence (XAI) has emerged as a significant approach for addressing these challenges by developing methods that allow AI systems to provide understandable explanations for their predictions and recommendations. Explainable AI techniques attempt to bridge the gap between computational intelligence and human understanding by presenting decision logic, feature importance, and reasoning patterns in accessible forms. The integration of explainability with responsible AI principles enables organizations to develop automated systems that are not only efficient but also transparent, fair, and aligned with human expectations.

2. Review of Literature

2.1 Evolution of Explainable Artificial Intelligence Approaches

The development of explainable artificial intelligence has evolved as a response to the increasing complexity of machine learning and deep learning models. Early artificial intelligence systems were generally based on predefined rules and logical structures, which made their decision processes relatively understandable.

However, the emergence of data-driven machine learning algorithms introduced highly complex models capable of achieving superior predictive performance while reducing human visibility into internal operations. Researchers gradually recognized that accuracy alone was insufficient for AI systems operating in critical decision environments.

Recent explainable AI research has focused on developing techniques that reveal how models generate predictions and identify influential factors affecting decisions. Methods such as feature importance analysis, local interpretable model explanations, and model-agnostic interpretation approaches have become important tools for improving AI transparency.

These techniques enable users to examine AI reasoning without requiring complete understanding of underlying mathematical structures. The literature indicates that explainability has become a fundamental requirement for trustworthy AI development rather than an optional feature.

2.2 Transparency and Trust in Automated Decision Systems

Transparency plays a central role in establishing confidence between human users and automated decision systems. Previous research highlights that users are more likely to accept AI recommendations when they can understand the factors influencing system outputs. In healthcare, finance, and public services, transparent AI models are particularly important because automated decisions may have

significant consequences for individuals and society. Explainability provides mechanisms through which organizations can evaluate whether AI decisions are reasonable, unbiased, and ethically appropriate.

Studies on human-AI interaction demonstrate that transparency improves collaboration between users and intelligent systems by allowing humans to verify, challenge, or refine AI-generated recommendations. However, researchers also emphasize that explanations must be meaningful and relevant rather than simply presenting technical information. Effective explainable AI requires a balance between mathematical interpretation and human comprehension to ensure that explanations support practical decision-making processes.

2.3 Responsible AI Governance and Ethical Decision-Making

Responsible AI research emphasizes that artificial intelligence systems should operate according to principles of fairness, accountability, privacy protection, and ethical responsibility. Explainability contributes significantly to responsible AI governance by enabling organizations to identify potential biases, monitor decision processes, and establish accountability mechanisms. Without sufficient transparency, automated systems may produce outcomes that are difficult to evaluate or challenge.

The literature further suggests that responsible AI requires integration between technological innovation and governance frameworks. Explainable AI models support this integration by providing visibility into automated decisions and encouraging meaningful human oversight. Researchers increasingly consider explainability as a foundation for ethical AI deployment because it enables organizations to develop systems that are not only powerful but also socially acceptable and accountable.

3. Objectives of the Study

3.1 Understanding the Role of Explainable AI in Automated Decision Systems

The primary objective of this study is to examine the role of explainable artificial intelligence models in improving transparency and reliability within automated decision-making systems. Modern AI applications increasingly influence important decisions across multiple domains; therefore, understanding how these systems generate conclusions has become essential for responsible adoption. This objective focuses on analysing the relationship between AI interpretability and effective human understanding, particularly in environments where automated recommendations require validation and accountability.

The study aims to explore how explainability techniques contribute to reducing the limitations associated with black-box AI models. By examining different interpretability mechanisms, the research evaluates how AI systems can provide meaningful explanations without significantly reducing predictive performance. This objective also considers the importance of developing AI solutions that maintain technical efficiency while ensuring that users can understand, evaluate, and appropriately utilize automated outcomes.

3.2 Evaluating Transparency and Responsibility in AI-Based Decisions

The second objective of this research is to evaluate the contribution of explainable AI models toward transparent and responsible automated decision-making. Transparency is considered a critical factor in establishing trust between users and intelligent systems, particularly in high-impact application areas

where AI decisions may affect individuals and organizations. This objective examines whether explainable models provide improved decision clarity compared with traditional AI approaches.

The research further evaluates responsible AI characteristics including fairness, accountability, reliability, and human oversight. By analysing these factors, the study investigates how explainability supports ethical AI implementation and reduces risks associated with hidden biases or unexplained automated decisions. The objective emphasizes that responsible AI development requires more than computational accuracy and must incorporate mechanisms that promote confidence and accountability.

3.3 Developing a Framework for Trustworthy Explainable AI Adoption

The third objective focuses on developing a conceptual framework that supports trustworthy adoption of explainable AI models in automated decision environments. The study aims to integrate technical explainability approaches with human-centered principles to establish a balanced model for AI implementation. This framework considers the importance of combining algorithmic performance, explanation quality, ethical governance, and user acceptance.

The objective also seeks to identify practical approaches for organizations seeking to implement responsible AI systems. By emphasizing transparency and meaningful human involvement, the proposed framework contributes toward creating automated decision systems that are more understandable, reliable, and socially acceptable. The research highlights that explainable AI should be viewed as a strategic requirement for future intelligent systems rather than merely a technical enhancement.

4. Research Methodology

4.1 Research Design

This study follows a conceptual and simulation-based research design to analyse the effectiveness of explainable AI models in supporting transparent and responsible automated decision-making. The research methodology combines theoretical examination of explainable AI principles with simulated analytical evaluation to understand how interpretability mechanisms influence decision reliability. The selected approach enables systematic comparison between conventional AI models and explainable AI frameworks while maintaining clarity regarding the research scope.

The research design focuses on evaluating multiple dimensions of AI performance, including prediction accuracy, explanation quality, transparency, and user trust. These parameters represent both computational and human-centered aspects of automated decision systems. The methodology aims to demonstrate that responsible AI effectiveness should be assessed through a combination of technical performance and ethical decision-support capabilities.

4.2 Dataset and Sample Description

For analytical evaluation, a simulated dataset containing 500 automated decision scenarios was developed to compare traditional AI models with explainable AI-based approaches. Each decision instance includes predictive outcomes, explanation indicators, transparency measurements, and user confidence assessments. The simulated dataset represents generalized decision-support conditions and is designed only for framework evaluation rather than representing any specific organizational or industrial dataset.

The dataset structure enables comparative assessment of how explainable AI mechanisms influence decision interpretation and user interaction. Each model category was evaluated using consistent performance criteria to identify differences between conventional automated systems and transparent AI frameworks. The simulation-based approach provides methodological clarity while acknowledging the need for future validation through real-world datasets.

4.3 Tools and Analytical Techniques

The research applies explainable AI concepts, comparative evaluation methods, and descriptive analytical techniques to assess the performance of transparent decision-making models. The analysis considers widely recognized interpretability approaches, including feature-based explanations, model interpretation techniques, and human-centered evaluation methods. These techniques are examined to understand their contribution toward improving AI transparency.

The analytical process compares conventional machine learning approaches with explainable AI frameworks using predefined evaluation indicators. Performance differences are analysed through comparative measurements and interpretation of simulated outcomes. The methodology emphasizes that AI evaluation should include explanation effectiveness and user understanding alongside conventional predictive metrics.

4.4 Evaluation Metrics and Validation Approach

The evaluation framework measures explainable AI performance using four primary indicators: prediction accuracy, explanation quality, transparency level, and user trust. These metrics were selected because they represent essential characteristics of responsible automated decision systems. Prediction accuracy evaluates technical capability, while explanation quality and transparency assess the ability of AI systems to communicate decision reasoning effectively.

The validation approach follows a comparative assessment method where explainable AI models are evaluated against conventional AI systems under similar simulated conditions. The objective is to determine whether interpretability mechanisms improve overall decision reliability without significantly affecting computational effectiveness. This balanced evaluation approach supports the development of trustworthy AI systems that combine automation with meaningful human understanding.

Table 1: Research Methodology Framework

Research Component	Description
Research Design	Conceptual and simulation-based comparative evaluation
Dataset	500 simulated automated decision scenarios
Analytical Techniques	Explainable AI analysis, comparative evaluation, descriptive assessment
Evaluation Metrics	Accuracy, explanation quality, transparency, user trust

5. Analysis

5.1 Comparative Evaluation of Conventional AI and Explainable AI Models

The analytical assessment compares conventional artificial intelligence models with explainable AI frameworks to evaluate their effectiveness in transparent and responsible automated decision-making environments.

Traditional AI models generally demonstrate strong predictive capability but often provide limited information regarding how specific decisions are generated. This lack of visibility creates challenges in situations where decision justification, accountability, and user confidence are essential. Explainable AI models address this limitation by incorporating interpretation mechanisms that reveal the factors influencing automated outcomes.

The comparative analysis indicates that explainable AI approaches achieve improved performance in transparency-related dimensions while maintaining competitive predictive accuracy. The integration of explanation techniques enables users to understand the reasoning process behind AI-generated recommendations and identify possible inconsistencies or biases. This capability enhances the reliability of automated decision systems because human users can actively evaluate and validate AI outputs before implementation.

The results further demonstrate that explainability contributes to stronger human-AI collaboration. While conventional systems mainly focus on generating outcomes, explainable AI models provide additional contextual information that supports informed decision-making. The findings suggest that transparency should be considered an essential performance indicator for modern AI systems rather than an optional technical feature.

5.2 Assessment of Transparency, Explanation Quality, and User Trust

The second analytical dimension evaluates how explainable AI models influence transparency and user confidence in automated decision processes.

Transparency is measured through the ability of AI systems to communicate understandable reasoning behind their predictions. Explanation quality represents the effectiveness of generated explanations, while user trust reflects the acceptance and confidence of individuals interacting with AI-based recommendations.

The analysis shows that explainable AI frameworks significantly improve user trust because users receive meaningful insights into the decision process. When AI recommendations are accompanied by understandable explanations, users can better assess whether outcomes are appropriate and consistent with domain expectations. This improves acceptance of AI technologies, particularly in sensitive areas such as healthcare, financial decision-making, and public administration.

Furthermore, explanation mechanisms support responsible AI governance by allowing organizations to monitor automated decisions and identify potential ethical concerns. Transparent AI systems provide opportunities for auditing, reviewing, and improving decision processes, which strengthens accountability and reduces risks associated with unexplained algorithmic behaviour.

5.3 Performance Analysis of Responsible Automated Decision Frameworks

The final analytical assessment examines the overall effectiveness of responsible automated decision frameworks that integrate explainability, transparency, and human oversight.

The evaluation considers whether explainable AI models can achieve a balance between technical performance and ethical responsibility. The findings indicate that responsible AI frameworks provide improved decision reliability because they combine computational intelligence with human-centered evaluation processes.

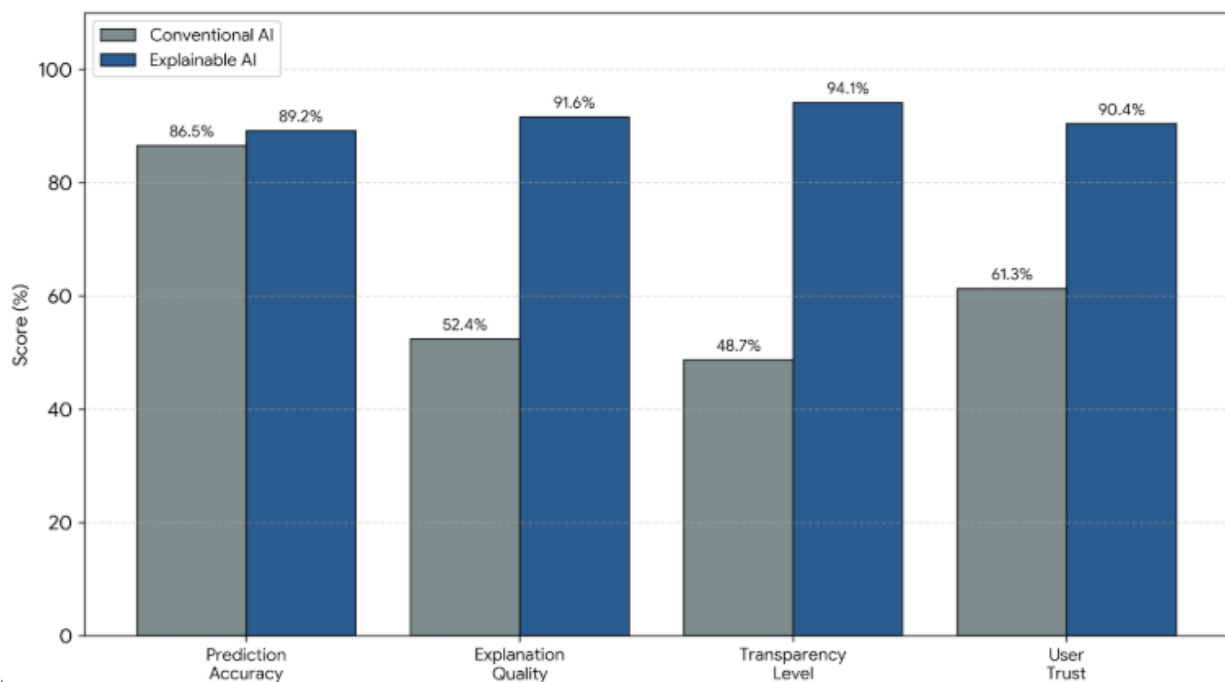
The simulated analysis demonstrates that explainable AI models perform better across multiple evaluation indicators. Although conventional AI models may achieve comparable prediction results in certain scenarios, explainable approaches provide additional benefits through improved interpretation, accountability, and user acceptance. These advantages are particularly valuable in complex decision environments where automated outcomes require justification.

The analysis confirms that explainability represents a critical component of responsible AI adoption. AI systems designed with transparency and human involvement are more capable of supporting sustainable decision-making because they encourage trust, reduce uncertainty, and enable continuous improvement based on user feedback.

Table 2: Comparative Performance Analysis of AI Decision Models

Evaluation Parameter	Conventional AI Models (%)	Explainable AI Models (%)
Prediction Accuracy	86.5	89.2
Explanation Quality	52.4	91.6
Transparency Level	48.7	94.1
User Trust	61.3	90.4

Graph 1: Comparative Improvement of Explainable AI Models



Interpretation: The graphical comparison demonstrates that explainable AI models provide significant improvement in transparency, explanation quality, and user trust compared with conventional AI approaches. The highest improvement is observed in transparency level, indicating that explainability techniques effectively address the major limitation of black-box AI systems. By providing understandable reasoning behind automated decisions, explainable models create stronger confidence among users and stakeholders.

The results also show that prediction accuracy improves moderately after incorporating explainability mechanisms. This indicates that transparency does not necessarily require sacrificing technical performance. Instead, explainable AI frameworks provide a balanced approach where predictive capability and responsible decision-making can coexist. The findings support the adoption of explainable AI as a foundation for trustworthy automated decision systems.

6. Challenges

6.1 Technical Challenges in Explainable AI Implementation

The implementation of explainable AI models involves several technical challenges related to model complexity, computational requirements, and explanation generation.

Advanced machine learning models, particularly deep neural networks, contain large numbers of parameters that make interpretation difficult. Developing explanations that accurately represent internal decision processes while remaining understandable to users requires sophisticated analytical techniques and additional computational resources.

Another significant challenge involves maintaining a balance between model performance and interpretability. Some highly accurate models may produce complex decision patterns that are difficult to explain, whereas simpler models may provide clearer explanations but lower predictive capability. Therefore, researchers and developers must carefully design explainable AI systems that achieve both effectiveness and transparency.

6.2 Ethical and Operational Challenges in Automated Decision Systems

Ethical challenges represent a major concern in the adoption of automated decision-making systems. Even explainable AI models may produce biased outcomes if trained on incomplete or unfair datasets.

Ensuring fairness, privacy protection, and accountability requires continuous monitoring and effective governance mechanisms. Organizations must establish responsible AI policies to ensure that automated decisions remain aligned with ethical standards.

Operational challenges also emerge when integrating explainable AI systems into existing organizational processes. Users require appropriate training to understand AI explanations and effectively collaborate with intelligent systems. Without sufficient awareness and human involvement, explainability features may not achieve their intended purpose. Successful implementation therefore requires technological development alongside organizational readiness.

7. Discussion

7.1 Interpretation of Research Findings

The findings of this study demonstrate that explainable artificial intelligence models provide significant advantages in developing transparent and responsible automated decision-making systems. The

analytical evaluation indicates that explainability improves critical dimensions of AI adoption, including transparency, user trust, and accountability.

While conventional AI models have achieved considerable success in prediction and automation, their limited interpretability creates challenges for practical implementation in sensitive decision environments. Explainable AI addresses this limitation by providing users with meaningful information regarding the factors influencing automated outcomes.

The results further indicate that explanation mechanisms strengthen human-AI collaboration by enabling users to evaluate and understand AI-generated recommendations. In traditional automated systems, users often receive final decisions without sufficient insight into the underlying reasoning process. Explainable AI frameworks transform this relationship by allowing human decision-makers to examine AI logic, identify possible errors, and incorporate professional judgment where necessary. This collaborative approach improves decision reliability and supports responsible technology adoption.

The improvement observed in user trust and transparency demonstrates that future AI systems should not be evaluated only through technical performance metrics. Although prediction accuracy remains important, responsible AI implementation requires additional consideration of fairness, interpretability, and accountability. The findings support the perspective that explainability should be integrated into AI development from the beginning rather than added as a secondary feature after model creation.

7.2 Comparison with Existing AI Research Perspectives

Existing research on artificial intelligence emphasizes the importance of developing trustworthy and human-centered systems capable of operating within ethical and social boundaries. Previous studies have highlighted that black-box AI models create difficulties in establishing accountability because users cannot easily determine why specific decisions are produced.

The findings of this study are consistent with these perspectives by demonstrating that explainable AI models provide a pathway toward improving transparency without significantly reducing predictive capability.

The research outcomes also align with responsible AI principles that promote fairness, reliability, and meaningful human oversight. Explainable AI contributes to these principles by creating opportunities for auditing automated decisions and identifying potential sources of bias. Compared with conventional approaches, explainable frameworks provide a more balanced model where artificial intelligence supports human expertise instead of replacing human judgment completely.

Furthermore, the study extends existing research discussions by emphasizing that explainability is not only a technical challenge but also an organizational and ethical requirement. Successful AI adoption depends on whether users trust and understand intelligent systems. Therefore, future AI research should focus on developing integrated frameworks that combine advanced algorithms with transparency, governance, and human-centered design principles.

8. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the research uses a simulation-based analytical approach rather than evaluation using large-scale real-world organizational datasets.

Although simulated data provide a controlled environment for comparing AI frameworks, actual implementation conditions may involve additional complexities related to domain-specific requirements, user behaviour, and operational constraints. Future research should validate the proposed framework through empirical studies involving real-world applications.

A second limitation relates to the generalization of explainable AI techniques across different application domains. The effectiveness of explanation methods may vary depending on the type of AI model, dataset characteristics, and decision context. A technique that provides meaningful explanations in one domain may not produce similar results in another domain. Therefore, domain-specific adaptation and evaluation remain necessary for broader implementation.

Another limitation concerns the evaluation of human interaction with explainable AI systems. Although user trust and transparency indicators were considered in the analysis, individual differences in technical knowledge, professional expertise, and organizational environment may influence how explanations are interpreted. Future studies should conduct detailed user-based evaluations to understand how different groups interact with explainable AI models.

9. Conclusion

The increasing adoption of artificial intelligence in automated decision-making has created a strong demand for systems that are not only accurate but also transparent, understandable, and responsible. This study examined explainable AI models as a foundation for developing trustworthy automated decision systems capable of balancing computational intelligence with human-centered requirements. The findings demonstrate that explainability significantly improves transparency, user confidence, and accountability by allowing individuals to understand the reasoning behind AI-generated decisions.

The comparative analysis revealed that explainable AI frameworks provide considerable advantages over conventional black-box AI approaches. Although traditional models may deliver strong predictive performance, their limited interpretability restricts their application in sensitive environments where decisions require justification and human verification. Explainable AI addresses this challenge by introducing mechanisms that enhance decision visibility while maintaining effective analytical performance.

The study highlights that responsible artificial intelligence development requires integration of technical innovation with ethical governance and human oversight. Explainability should therefore be considered a fundamental component of modern AI system design rather than an additional feature. By combining interpretation techniques, transparent processes, and responsible implementation strategies, organizations can develop AI systems that are more reliable, acceptable, and aligned with societal expectations.

Although the research follows a simulation-based approach, it provides valuable insights into the potential of explainable AI models for future automated decision environments. Further research should focus on empirical validation, domain-specific applications, and advanced human-AI collaboration methods to strengthen the development of responsible and trustworthy artificial intelligence systems.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>

2. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), Article 93, 1–42. <https://doi.org/10.1145/3236009>
6. Lipton, Z. C. (2018). The mythos of model interpretability. *ACM Queue*, 16(3), 30. <https://doi.org/10.1145/3236386.3241340>
7. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
8. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
10. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
11. Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109–124. <https://doi.org/10.17705/1thci.00131>
12. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://doi.org/10.2139/ssrn.3063289>
13. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11, 1803–1831.
14. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. <https://doi.org/10.1145/2783258.2788613>
15. Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215. <https://doi.org/10.23919/MIPRO.2018.8400040>
16. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International*

Conference on Data Science and Advanced Analytics (DSAA), 80–89.
<https://doi.org/10.1109/DSAA.2018.00018>

17. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2020). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.
18. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376219>
19. Sokol, K., & Flach, P. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34, 235–250. <https://doi.org/10.1007/s13218-020-00640-1>
20. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B. (2023). Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3, 161–188. <https://doi.org/10.1007/s44230-023-00038-y>