

Explainable AI Frameworks for Improving Transparency in Automated Decision-Making Systems

Dr. Emma van Leeuwen

Assistant Professor, Utrecht Institute of Advanced Computing, Netherlands

Abstract

The increasing adoption of artificial intelligence in automated decision-making has created a growing need for systems capable of providing understandable and verifiable explanations for their outputs. Machine learning algorithms can process complex information and generate predictions across diverse application domains, yet sophisticated predictive models often provide limited insight into the factors influencing individual decisions. This lack of transparency can complicate accountability, reduce opportunities for meaningful human oversight, and make it difficult to identify inappropriate dependencies within automated processes. Explainable artificial intelligence offers methodological approaches for addressing these concerns through interpretable models, feature-attribution techniques, local explanations, and structured decision documentation. This hypothetical methodological study examines explainable AI frameworks for improving transparency in automated decision-making systems. The proposed approach integrates interpretable surrogate modeling, explanation fidelity assessment, explanation stability evaluation, and human-centered interpretation. A hypothetical analytical scenario is constructed to illustrate the relationship between explanation complexity, agreement with an underlying predictive model, and consistency under small input changes. Two illustrative comparisons are developed using assumed explanation-agreement counts and perturbation-stability outcomes.

The analytical calculations demonstrate how explanatory methods can differ across fidelity and stability dimensions. More detailed explanations are assigned greater agreement with the illustrative predictive system, while simpler explanations are assigned greater stability under the specified assumptions. These relationships are features of the hypothetical scenario rather than experimentally established findings. The study contributes a structured methodological framework for evaluating explainable AI systems beyond the mere availability of explanatory outputs. It emphasizes that transparency requires appropriate explanation fidelity, understandable presentation, documented limitations, and mechanisms for responsible review. The paper concludes that explainable AI can support more accountable automated decision-making, but reliable implementation requires empirical validation and careful consideration of explanation accuracy, usability, fairness, and system governance.

Keywords: Explainable Artificial Intelligence, Automated Decision-Making, Algorithmic Transparency, Interpretable Machine Learning, Explanation Fidelity, Explainability, Decision Accountability, Model Interpretability, Human Oversight, Artificial Intelligence Governance.

1. Introduction

Artificial intelligence has become an important component of contemporary information systems because it enables computational models to process complex datasets and support decisions that previously required extensive human assessment. Automated decision-making systems are increasingly associated with financial services, healthcare management, organizational operations, and digital platforms. Machine learning algorithms can identify nonlinear relationships and generate predictions from large collections of information. However, the technical sophistication of these systems does not necessarily provide an understandable account of how individual decisions are produced. This distinction has increased scholarly interest in transparency and interpretability.

Complex predictive algorithms often contain numerous parameters, interacting variables, and computational transformations that are difficult to interpret directly. Ensemble methods, nonlinear classifiers, and deep learning architectures may produce useful predictions while providing limited visibility into the relationships supporting their outputs. Breiman (2001) and Friedman (2001) established influential predictive learning approaches that demonstrate the capabilities of complex statistical modeling. Nevertheless, predictive performance alone cannot determine whether a decision-making system is sufficiently understandable for its intended application. Meaningful evaluation must also consider whether relevant stakeholders can examine the factors influencing automated recommendations.

Explainable artificial intelligence addresses this methodological problem by developing procedures that make selected aspects of computational decision processes more accessible. Such procedures include feature contribution analysis, simplified local models, interpretable rules, and visual representations of model behavior. Štrumbelj and Kononenko (2010) investigated explanations based on feature contributions, while Lou et al. (2013) examined intelligible predictive models incorporating selected feature interactions. Ribeiro et al. (2016) subsequently developed a model-agnostic approach for explaining individual classifications. These contributions demonstrated that explanation methods can support the examination of complex predictive behavior without necessarily requiring complete access to a model's internal structure.

Transparency also involves considerations extending beyond mathematical interpretation. Automated decisions may influence individuals and institutions, making accountability, fairness, and the ability to identify unsuitable model behavior important concerns. Barocas and Selbst (2016) examined how data-driven decisions may reproduce or reinforce discriminatory patterns, while Mittelstadt et al. (2016) analyzed ethical questions surrounding algorithmic systems. Their contributions suggest that explanatory information should be evaluated in relation to the broader conditions under which automated decisions are made. An explanation that appears understandable may still be incomplete, misleading, or insufficient for establishing responsible decision-making practices.

The present study examines explainable AI frameworks through a hypothetical methodological analysis of automated classification decisions. It focuses on the relationship between explanation complexity, fidelity to an underlying predictive model, and stability under small changes in input information. A simplified analytical scenario is used to demonstrate alternative explanation configurations without claiming actual algorithmic implementation. The research also considers limitations associated with explanation reliability, information quality, user comprehension, and governance. Its purpose is to establish a structured foundation for evaluating transparency-oriented AI frameworks and identifying requirements for subsequent empirical validation.

2. Objectives of the Study

The primary objective of this study is to examine how explainable artificial intelligence frameworks may improve transparency in automated decision-making systems. The investigation focuses on interpreting predictive outcomes, evaluating explanation reliability, and identifying implementation challenges. The objectives establish a methodological foundation for assessing transparency through hypothetical analytical comparisons.

2.1 Examination of Explainable AI Techniques

The first objective is to examine techniques that make automated decision-making processes more understandable to users and reviewers. Particular attention is directed toward interpretable rule models, feature-attribution methods, local surrogate explanations, and visual representations of predictive behavior. These approaches provide different forms of information about the relationships influencing model outputs. The study considers their methodological characteristics and investigates how explanation structure may affect the information available for decision review. This examination establishes a conceptual foundation for evaluating alternative transparency-oriented learning frameworks.

The objective also involves distinguishing explanation availability from explanation reliability. A computational system may generate detailed statements about its predictions without demonstrating that those statements accurately represent the underlying model. Consequently, explanatory methods must be assessed according to their connection with actual predictive behavior. The study considers local fidelity, interpretive complexity, and consistency as relevant evaluation dimensions. Understanding these distinctions supports the development of frameworks that prioritize faithful communication of model behavior rather than producing explanations solely because they appear understandable or persuasive.

2.2 Evaluation of Explanation Transparency

The second objective is to evaluate how explanation complexity may influence the transparency of automated decisions under a defined hypothetical scenario. The study considers alternative surrogate explanations containing different numbers of interpretable conditions. Illustrative agreement counts are used to demonstrate how closely simplified explanations might reproduce the predictions of an underlying classification system. These values are mathematical assumptions and do not represent outcomes obtained from implemented explanation algorithms. The objective is to establish a transparent method for comparing explanatory configurations.

Another purpose is to examine explanation stability under small variations in input information. Stable explanations may support consistent interpretation when nearby observations receive similar predictions, although meaningful model changes can justify different explanations. The study introduces hypothetical perturbation pairs to illustrate alternative stability outcomes. This assessment distinguishes predictive agreement from explanation consistency because the two properties describe different characteristics of an explanatory method. The objective therefore promotes multidimensional evaluation rather than treating a single transparency measure as sufficient evidence of explanatory quality.

2.3 Identification of Accountability Challenges

The third objective is to identify challenges associated with interpreting and governing automated decision-making systems. These challenges include model complexity, incomplete information, unstable

explanations, inappropriate feature dependencies, and differences in user expertise. The study examines how such conditions may affect the reliability of explanatory information and complicate the identification of decision errors. Understanding these limitations contributes to designing evaluation procedures that distinguish technical explanation quality from broader questions concerning fairness, accountability, and responsible organizational decision-making.

The investigation also considers the importance of maintaining verifiable records of automated decisions and their associated explanations. Explanatory outputs should be linked to identifiable model versions, input conditions, and established evaluation procedures. Such documentation may support systematic review and help distinguish model behavior from subsequent interpretation. The objective emphasizes that transparency cannot be established through explanatory text alone. Reliable accountability requires appropriate technical evidence, meaningful human oversight, and clear recognition of the limitations associated with automated decision-support systems.

3. Review of Literature

3.1 Development of Interpretable Machine Learning

Breiman (2001) established random forests as an influential ensemble learning approach, while Friedman (2001) developed gradient boosting through statistical optimization. These methods expanded predictive modeling capabilities but also illustrated the difficulty of understanding complex interactions within sophisticated learning systems. Friedman and Popescu (2008) subsequently examined predictive rule ensembles, demonstrating how combinations of interpretable rules can support both classification and model understanding. Their research established an important connection between predictive learning and the construction of explanations grounded in recognizable decision conditions.

Lou et al. (2013) investigated intelligible models incorporating selected pairwise interactions, demonstrating that interpretable model structures can represent relationships more complex than purely additive effects. Caruana et al. (2015) further examined interpretable predictive models in healthcare applications, emphasizing the importance of understanding model behavior in consequential settings. Letham et al. (2015) developed interpretable classifiers based on rules and Bayesian analysis. Together, these studies demonstrate that transparency can be incorporated directly into model design rather than treated exclusively as an additional explanatory procedure.

3.2 Local Explanations and Feature Contributions

Štrumbelj and Kononenko (2010) examined a game-theoretic approach to explaining individual classifications through feature contributions. Their subsequent research extended feature-based explanation methods to broader predictive models (Štrumbelj & Kononenko, 2014). Goldstein et al. (2015) developed individual conditional expectation plots for examining heterogeneous predictive relationships. These contributions established methods for investigating how input variables influence model outputs, while also demonstrating that understanding an individual prediction may require analysis different from that used to characterize global model behavior.

Bach et al. (2015) investigated layer-wise relevance propagation for explaining nonlinear classification decisions, providing a framework for attributing model outputs to input components. Ribeiro et al. (2016) introduced local interpretable model-agnostic explanations, demonstrating how simplified models can approximate a classifier within a selected neighborhood. Datta et al. (2016) examined quantitative input influence as a foundation for algorithmic transparency. These studies

provide distinct approaches to explanatory analysis and highlight the importance of specifying exactly what an explanation represents and how its reliability should be evaluated.

3.3 Human Understanding and Decision Accountability

Lim et al. (2009) investigated explanation types in context-aware intelligent systems and examined how explanations influence user understanding. Kulesza et al. (2015) studied explanatory debugging as a means of helping users understand and correct learning-system behavior. Their contributions demonstrate that explanations should be assessed in relation to the needs and capabilities of intended users. An explanation that communicates useful information to a technical specialist may not provide sufficient clarity for individuals without machine learning expertise.

Dwork et al. (2012) developed a theoretical approach to fairness in algorithmic decision-making, while Zemel et al. (2013) examined fair representations in machine learning. Hardt et al. (2016) investigated equality of opportunity in supervised prediction. Barocas and Selbst (2016) and Mittelstadt et al. (2016) further examined legal and ethical concerns associated with algorithmic systems. These studies establish that transparency should be considered alongside fairness and accountability, although an understandable explanation does not independently demonstrate that a decision-making system is fair.

4. Research Methodology

4.1 Research Design and Explainability Framework

The present study adopts a hypothetical methodological research design supported by illustrative analytical calculations. Its proposed framework considers a binary automated classification system whose outputs are examined through simplified local explanations. Three hypothetical explanation configurations are developed to represent different levels of interpretive complexity. Each configuration is characterized by a specified maximum number of explanatory rule conditions. The framework evaluates how closely these explanations reproduce the reference model's predictions and how consistently their explanatory conditions are retained under limited input variation.

The methodological framework distinguishes prediction, explanation generation, and explanation evaluation as separate activities. A reference classifier first produces a decision using available input information. A surrogate explanation then approximates the decision behavior within a defined local neighborhood. Finally, fidelity and stability measures are used to assess selected characteristics of the explanation. This separation is important because the correctness of a prediction does not establish the faithfulness of its explanation. No reference classifier or surrogate model has actually been trained for this illustrative study.

4.2 Hypothetical Data and Explanation Configurations

The illustrative evaluation assumes 200 local assessment instances for each explanation configuration. These instances represent conceptual observations selected from a hypothetical automated decision-making environment. Three surrogate explanations are considered, containing a maximum of two, four, and eight rule conditions. The configurations are designated as concise, intermediate, and detailed. Each is assigned an illustrative number of predictions agreeing with the reference classifier. The assigned counts are not derived from experimental observations and are intended solely to demonstrate a reproducible performance calculation.

A separate stability assessment assumes 100 matched pairs of closely related input observations for each configuration. Within each pair, small non-decisive changes are assumed not to alter the reference classifier's predicted decision. Explanation stability is defined as retention of the same explanatory condition set across the pair. Hypothetical stability counts are assigned to demonstrate differences among the configurations. The assumption that perturbations preserve the reference decision is important because explanations may reasonably change when underlying model behavior changes substantially.

4.3 Analytical Procedure and Evaluation Measures

The first analytical procedure calculates local explanation fidelity using assumed agreement counts. Fidelity is defined as the proportion of neighborhood observations for which a surrogate explanation reproduces the reference classifier's predicted label. If A represents agreement count and N represents the total number of assessed observations, the measure is expressed as $\text{Fidelity} = (A/N) \times 100$. This definition concerns agreement with model outputs rather than correctness against verified ground-truth labels. A faithful explanation can therefore describe an inaccurate or otherwise problematic predictive model.

The second procedure calculates explanation stability from hypothetical perturbation pairs. If S represents the number of pairs retaining identical explanatory condition sets and P represents the total number of eligible pairs, stability is expressed as $\text{Stability} = (S/P) \times 100$. The analysis compares these measures descriptively without treating them as experimental estimates. In future implementations, explanation complexity should additionally be evaluated through user-comprehension assessments, runtime measurements, and tests of whether explanatory claims remain consistent with the underlying model.

5. Analysis

5.1 Local Explanation Fidelity Across Model Configurations

The first hypothetical analysis considers the relationship between explanation complexity and agreement with an underlying automated classifier. Each explanation configuration is evaluated against 200 assumed local observations. The concise configuration contains a maximum of two explanatory conditions, whereas the intermediate and detailed configurations contain four and eight conditions respectively. The hypothetical agreement counts increase across these configurations. This arrangement illustrates how explanation fidelity can be represented quantitatively without claiming that more complex explanations necessarily perform better in real applications.

The illustrative calculations assign 168 agreements to the concise configuration, 182 to the intermediate configuration, and 192 to the detailed configuration. Their corresponding fidelity values are 84%, 91%, and 96%. These quantities demonstrate the mathematical relationship between agreement counts and fidelity percentages. They do not establish that an eight-condition explanation is inherently superior to a simpler alternative. Actual fidelity depends on the reference model, the local neighborhood, the explanation algorithm, and the selected evaluation procedure.

Table 1: Hypothetical Local Explanation Fidelity Across Three Configurations

Explanation Configuration	Maximum Rule Conditions	Agreement Count	Disagreement Count	Local Fidelity (%)
Concise	2	168	32	84
Intermediate	4	182	18	91
Detailed	8	192	8	96

Note: Each configuration is assessed against 200 assumed local observations. Counts and percentages are illustrative and do not represent the performance of implemented explainable AI algorithms.

5.2 Explanation Stability Under Input Perturbations

The second hypothetical analysis examines explanation consistency under small changes in input information. Each explanation configuration is assessed using 100 conceptual perturbation pairs in which the underlying reference decision is assumed to remain unchanged. Stability is measured through retention of the same explanatory condition set within each pair. The assigned outcomes differ across the three configurations, providing an illustrative comparison between explanatory complexity and condition-set consistency. These values are assumptions rather than observed results from a perturbation experiment.

The concise configuration retains its explanatory conditions across 90 assumed pairs, producing stability of 90%. The intermediate configuration retains identical conditions across 78 pairs, while the detailed configuration does so across 62 pairs. These illustrative results demonstrate that fidelity and stability can be assessed independently. The assigned pattern should not be interpreted as a universal complexity–stability relationship. In real systems, explanation stability depends on model structure, feature dependence, perturbation design, and the meaning of differences between explanatory outputs.

Table 2: Hypothetical Explanation Stability Across Perturbation Pairs

Explanation Configuration	Perturbation Pairs	Stable Explanations	Changed Explanations	Stability (%)
Concise	100	90	10	90
Intermediate	100	78	22	78
Detailed	100	62	38	62

Note: Stability represents the assumed proportion of pairs retaining identical explanatory condition sets. The values are illustrative and are not outputs from an experimental explanation system.

6. Discussion

6.1 Explainability and Algorithmic Transparency

The hypothetical findings demonstrate how explanation fidelity can be incorporated into a methodological assessment of automated decision-making transparency. An explanatory model may

communicate selected relationships in an understandable form, but its usefulness depends partly on how accurately it represents the predictive behavior being examined. Ribeiro et al. (2016) emphasized the value of local approximations for interpreting complex classifiers. The present analytical framework extends this perspective by treating fidelity as a measurable property that requires explicit assessment rather than assumption.

Transparency also depends on the limitations communicated alongside an explanation. A local surrogate may represent decisions accurately within a selected neighborhood while failing to describe the reference model elsewhere. Consequently, explanatory statements should identify the observations and conditions to which they apply. The hypothetical analysis does not establish that an explanation faithfully identifies causal relationships or objectively correct decision criteria. Instead, it emphasizes that interpretable outputs should be connected to the underlying predictive system through reproducible verification procedures.

6.2 Explanation Complexity and Interpretive Stability

The analytical comparison illustrates that explanation complexity and stability can be treated as distinct evaluation dimensions. A detailed surrogate may reproduce more of the underlying model's local behavior while introducing additional conditions that could vary across nearby observations. However, this pattern is not inevitable. A well-designed detailed explanation may remain stable, whereas an oversimplified explanation may change unpredictably. The hypothetical values merely demonstrate how alternative configurations can be compared using clearly specified mathematical measures.

An important implication is that simplified explanations should not automatically be considered more reliable because they contain fewer conditions. Excessive simplification may conceal relevant feature interactions or produce misleading interpretations. Conversely, highly detailed explanations may become difficult for nontechnical users to understand. Research on intelligible models by Lou et al. (2013) and Caruana et al. (2015) demonstrates the importance of balancing predictive complexity with meaningful interpretation. Such considerations support multidimensional explanation assessment rather than reliance on a single numerical indicator.

6.3 Human Understanding and Decision Accountability

Explainable AI frameworks should be evaluated according to whether their outputs support meaningful examination of automated decisions. Technical explanations may improve access to model information without guaranteeing that users understand the underlying reasoning. Lim et al. (2009) and Kulesza et al. (2015) demonstrated the relevance of explanation design and user interaction. Their work suggests that explanatory systems should consider the knowledge, responsibilities, and informational needs of their intended audiences rather than presenting a uniform explanation format for every application.

Accountability also requires recognizing that technically faithful explanations do not establish fairness or justify consequential decisions. An explanation may accurately identify influential variables while revealing an undesirable dependence within the predictive model. Research by Barocas and Selbst (2016) and Mittelstadt et al. (2016) highlights broader concerns surrounding automated decision systems. Consequently, explanation quality should be considered alongside independent evaluation of predictive validity, fairness, decision procedures, and the availability of appropriate human review mechanisms.

7. Challenges

7.1 Model Complexity and Explanation Reliability

Complex automated decision-making systems may incorporate nonlinear interactions, high-dimensional features, and multiple computational transformations. These characteristics make it difficult to develop explanations that remain both understandable and faithful to the underlying model. Simplified local approximations may omit interactions that influence particular decisions, while highly detailed explanations can become difficult to interpret. The challenge is therefore not merely generating explanatory information but establishing whether that information accurately reflects the predictive behavior relevant to a specific decision.

Another difficulty concerns differences between explanatory techniques. Feature-attribution methods, rule-based explanations, and interpretable surrogate models communicate different aspects of computational behavior. Their outputs may not be directly comparable without a common evaluation framework. An explanation that identifies influential variables does not necessarily reveal how those variables interact or whether changing them would alter the final decision. Reliable evaluation consequently requires clearly defined explanation objectives, suitable reference conditions, and careful interpretation of methodological limitations.

7.2 Data Bias and Ethical Interpretation

Automated decision-making systems may reproduce patterns contained in historical training information, including relationships that create inappropriate differences among individuals or groups. Explainable AI can reveal some of these dependencies, but explanatory visibility does not automatically eliminate them. Dwork et al. (2012) and Hardt et al. (2016) investigated formal approaches to fairness, demonstrating that fairness involves criteria requiring explicit definition and assessment. Ethical evaluation therefore extends beyond identifying the variables associated with particular model outputs.

A related challenge concerns the interpretation of proxy variables and correlated features. Removing an explicitly sensitive attribute from a model does not necessarily prevent other variables from carrying related information. Explanations may identify such relationships without conclusively establishing whether a particular decision violates an appropriate fairness criterion. Accurate assessment requires attention to the data-generating process, intended application, relevant comparison groups, and consequences of model decisions. Transparent explanation should therefore support, rather than replace, independent fairness and accountability evaluations.

7.3 User Comprehension and Operational Governance

The effectiveness of an explanation depends partly on the knowledge and responsibilities of the individual receiving it. Technical specialists may require information about feature interactions and local model fidelity, while other users may need concise explanations identifying relevant decision factors and available review procedures. Providing identical technical outputs to every audience may reduce practical understanding. Explainable AI frameworks should therefore distinguish model-oriented diagnostics from explanations designed for operational decision review and communication.

Operational governance presents additional difficulties involving model updates, explanation consistency, documentation, and responsibility for automated outcomes. An explanation generated under one model version may no longer represent predictions produced after retraining. Systems should therefore preserve relevant information about model configurations, evaluation conditions, and

explanatory methods. Appropriate oversight also requires identifying who can question, review, or modify consequential decisions. These safeguards establish a stronger foundation for accountability than explanatory displays presented without supporting technical records or review mechanisms.

8. Conclusion

Explainable artificial intelligence offers methodological approaches for improving transparency in automated decision-making systems through interpretable models, local explanations, and feature-based analysis. The present hypothetical study examined explanatory reliability using a framework that distinguishes local fidelity, explanation complexity, and stability. Its discussion indicates that meaningful transparency requires more than the production of understandable statements. Explanations must be connected to actual model behavior and evaluated according to clearly specified technical and interpretive requirements.

The illustrative analytical comparisons demonstrated how different explanation configurations can be assessed through agreement counts and perturbation-stability measures. Under the assumed values, detailed explanations displayed higher local fidelity, whereas concise explanations showed greater stability. These relationships were constructed for mathematical demonstration and do not establish experimentally verified performance patterns. Their significance lies in showing that explanation quality has multiple dimensions that should be evaluated separately and interpreted within the conditions of a particular decision-making environment.

The study concludes that dependable explainable AI frameworks should integrate faithful model interpretation with understandable communication, responsible oversight, and transparent evaluation procedures. Future empirical investigations should compare explanation techniques using representative datasets, controlled perturbations, and appropriate user assessments. Particular attention should be directed toward explanation accuracy, fairness, stability, and the relationship between interpretive information and decision accountability. Such investigations can establish the practical conditions under which explainable AI meaningfully improves automated decision-making transparency.

References

1. Abusitta, A., Li, M. Q., & Fung, B. C. M. (2024). Survey on explainable AI: Techniques, challenges and open issues. *Expert Systems with Applications*, 255, Article 124710. <https://doi.org/10.1016/j.eswa.2024.124710>
2. 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>
3. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html>
4. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bannetot, 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>
5. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), Article 832. <https://doi.org/10.3390/electronics8080832>

6. Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615, 238–292. <https://doi.org/10.1016/j.ins.2022.10.013>
7. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://arxiv.org/abs/1702.08608>
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), Article 93. <https://doi.org/10.1145/3236009>
9. Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable artificial intelligence-based decision support systems: A recent review. *Electronics*, 13(14), Article 2842. <https://doi.org/10.3390/electronics13142842>
10. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, Article 103473. <https://doi.org/10.1016/j.artint.2021.103473>
11. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. <https://doi.org/10.1145/3233231>
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
13. Mersha, M., Lam, K., Wood, J., AlShami, A. K., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599, Article 128111. <https://doi.org/10.1016/j.neucom.2024.128111>
14. 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>
15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 1527–1535. <https://doi.org/10.1609/aaai.v32i1.11491>
17. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
18. Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, 14(19), Article 8884. <https://doi.org/10.3390/app14198884>
19. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), 247–278. <https://doi.org/10.1109/JPROC.2021.3060483>
20. Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2239–2250). Association for Computing Machinery. <https://doi.org/10.1145/3531146.3534639>