

CHAPTER 5

EXPLAINABLE ARTIFICIAL INTELLIGENCE: INTERPRETING BLACK-BOX MODELS AND CURRENT APPROACHES

İsmail AKGÜL¹

1. Introduction

Artificial intelligence (AI) systems can learn patterns from large amounts of data and perform classification, prediction, and decision-support tasks. Advances in machine learning (ML) and deep learning (DL), in particular, have created significant opportunities for solving complex problems. Nevertheless, the decision-making processes of some models with high predictive performance cannot be easily traced or understood by humans. Large numbers of parameters, nonlinear relationships, and interactions among features can make it difficult to explain why a particular prediction was produced (Arrieta et al., 2020). This issue is commonly referred to in the literature as the black-box problem (Guidotti et al., 2018). Although a model's inputs and outputs may be observable, the inability to understand directly the computational processes that lead to its outputs can make the model difficult to

¹ Assoc. Prof. Dr., Erzincan Binali Yıldırım University, Faculty of Engineering and Architecture, Department of Computer Engineering, Orcid: 0000-0003-2689-8675

evaluate. For example, when a credit assessment system rejects an application, simply communicating the decision is not sufficient to establish which features influenced it. Similarly, when a clinical decision-support system indicates a high risk of disease, questions arise about how the prediction should be interpreted in a clinical context.

Explainable artificial intelligence (XAI) encompasses the methods and approaches developed to address these needs. Such methods can serve different purposes, ranging from examining the features associated with an individual prediction to evaluating a model's overall behavior. Explainability is therefore a broad field that cannot be reduced to a single algorithm or a uniform form of explanation (Ding et al., 2022; Schwalbe & Finzel, 2024; Fridkin & Bendersky, 2026). However, a model's ability to generate an explanation does not necessarily mean that it is accurate or trustworthy. Incorrect predictions can also be presented in an explainable form. Moreover, an explanation that appears convincing to users may not accurately reflect the decision patterns actually used by the model. The sensitivity tests developed by Adebayo et al. (2018) to evaluate explanation methods are important because they demonstrate that judging explanations solely by how plausible they appear is insufficient.

This chapter examines the fundamental concepts and principal methods of XAI, compares the differences between these methods, and evaluates their contributions across application domains. Its central approach is to treat explainability not merely as the production of a technical output, but as part of a broader process of understanding model behavior, identifying limitations, and evaluating results in the appropriate context.

2. Fundamental Concepts of Explainable Artificial Intelligence (XAI)

2.1. Explainability and Interpretability

Although the terms explainability and interpretability are frequently used together in the literature, they can emphasize different aspects of understanding a model. Interpretability refers to the extent to which a model's structure or decisions can be understood directly by humans. Explainability, by contrast, focuses on providing rationales and representations that help users understand a model's predictions or behavior. Nevertheless, the literature does not establish a universally accepted, precise distinction between these concepts (Lipton, 2018; Arrieta et al., 2020). Linear models and small decision trees can serve as directly interpretable models under appropriate conditions. Their structures can be examined to understand the effects of individual variables or the rules governing decisions. By contrast, it is more difficult to trace the internal computations that produce predictions in large ensemble models and complex neural networks. Post-hoc explanation methods can be used to investigate the behavior of these models.

An important distinction here is between explaining a complex model after it has been developed and designing an interpretable model from the outset. Post-hoc methods attempt to represent an original model's individual predictions or overall behavior in an understandable form. However, the explanatory model and the model being explained are not the same. Therefore, the extent to which an explanation represents the original model's behavior must be evaluated separately (Ribeiro et al., 2016; Rudin, 2019; Di Marino et al., 2025). Directly interpretable models should also be considered in high-stakes applications. Model complexity does not necessarily lead to better results in every situation. Likewise, choosing an interpretable model does not automatically

guarantee more accurate decisions. Model selection should take predictive performance, explainability, the costs of errors, and the intended application context into account.

2.2. Local and Global Explanations

Explanations can be classified as local or global according to their scope. Local explanations focus on understanding the rationale behind a prediction for a specific instance. For example, they can be used to investigate why a particular credit application was rejected or why an image was assigned to a certain class. LIME and instance-level SHAP explanations can be used for this purpose (Ribeiro et al., 2016; Lundberg & Lee, 2017; Lin et al., 2026). Global explanations, on the other hand, aim to examine a model's overall behavior, the importance of variables across a dataset, or patterns that emerge across different inputs. Feature importance analyses and certain feature relationship plots can be used in such evaluations (Friedman, 2001; Molnar, 2022).

These two approaches address different questions. A feature that is generally important does not necessarily have the same effect on every individual prediction. Similarly, an explanation generated for a single instance may not represent the model's behavior across the entire dataset. Accordingly, using local and global explanations together may be appropriate when a comprehensive assessment of a model is required.

2.3. Model-Specific and Model-Agnostic Methods

Explanation methods can also be classified according to whether they require access to a model's internal structure. Model-agnostic methods generally generate explanations using the model's inputs and outputs. This makes it possible to apply the same approach to different types of models. LIME is one of the principal methods in this category (Ribeiro et al., 2016). Model-specific methods, by contrast, rely on the internal structure, parameters, or

gradients of particular model families. For example, Gradient-weighted Class Activation Mapping (Grad-CAM) was developed to visualize image regions associated with a particular class prediction in convolutional neural networks (CNNs) (Selvaraju et al., 2017; Fayyaz et al., 2025).

Although this classification can help guide method selection, it does not establish a ranking of explanation quality. A model-agnostic method is not necessarily more reliable in every situation, just as a method that uses a model's internal structure does not automatically produce more accurate explanations. The suitability of a method should be considered in relation to the problem being explained and the criteria used to evaluate the explanation.

2.4. Evaluating Explanations

The success of an explanation depends on more than its ability to produce an understandable graph or textual description. An explanation should accurately represent model behavior, avoid unnecessary changes under similar conditions, and provide useful information to its intended users. Various technical metrics and user-centered evaluation approaches are used to assess explanation quality (Carvalho et al., 2019; Naveed et al., 2024; Fridkin & Bendersky, 2026). The principal criteria include the following:

- **Fidelity:** The extent to which an explanation accurately represents the behavior of the model being explained.
- **Stability:** Whether explanations change unnecessarily for similar inputs.
- **Understandability:** The extent to which the intended user can comprehend the explanation.
- **Actionability:** The extent to which the explanation helps identify an appropriate next step.

- **Scope:** The instances or conditions to which the explanation applies.

These criteria are not independent of one another. For example, an easy-to-understand explanation may be misleading if it does not accurately represent model behavior. Conversely, an explanation that is technically faithful may be too complex for users to understand. The technical properties of an explanation and the benefits it provides to users must therefore be evaluated together.

3. Principal Methods Used in XAI

Explanation methods differ in the types of information they provide and the questions they are designed to answer. Some assign feature contributions to predictions, whereas others construct a simple model around a particular prediction. Counterfactual methods investigate which changes may be required to obtain alternative outcomes, while gradient-based methods examine the relationship between model outputs and inputs. Method selection should therefore be guided by the specific explanatory need (Ding et al., 2022; Schwalbe & Finzel, 2024).

3.1. Shapley Additive Explanations (SHAP)

SHAP is a family of methods developed to explain the contributions of individual features to ML predictions. Lundberg and Lee (2017) introduced this approach as a unified framework that brings different additive feature attribution methods together. Its theoretical foundation lies in Shapley values from cooperative game theory. The central idea is to assess a feature's contribution by considering its marginal contributions when added to different subsets of features.

The additive structure of SHAP can generally be expressed as shown in Equation 1:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z_i' \quad (1)$$

Here, g denotes the explanation model, ϕ_0 represents the reference prediction value, ϕ_i denotes the contribution attributed to the corresponding feature, and z_i' represents the feature in the explanation space. M is the number of features considered in the explanation. This formulation represents a prediction in terms of a reference value and the contributions attributed to individual features (Lundberg & Lee, 2017).

One advantage of SHAP is that it can be used not only to examine individual predictions but also to analyze multiple predictions collectively. For example, in a risk prediction model, it is possible to investigate which variables increase or decrease the score assigned to a particular application. When explanations for multiple applications are considered together, additional insights into the model's overall behavior can be obtained.

Nevertheless, SHAP values must be interpreted carefully. Results may depend on how features are handled and which reference data are used. For correlated variables, the assumptions adopted can also influence how contributions are distributed among features. Furthermore, a feature's substantial contribution to a prediction does not prove that the feature causes the outcome in the real world. SHAP is therefore a useful tool for examining predictive behavior, but it is not a substitute for causal inference (Molnar, 2022; Salih et al., 2025).

3.2. Local Interpretable Model-agnostic Explanations (LIME)

LIME aims to explain a specific prediction made by a complex model by constructing a simpler model around the instance of interest. Ribeiro et al. (2016) proposed generating new samples in the neighborhood of the instance to be explained, evaluating these

samples using the original model, and learning an interpretable model that approximates its behavior in the local region.

The basic approach underlying LIME can be expressed as shown in Equation 2:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} [L(f, g, \pi_x) + \Omega(g)] \quad (2)$$

Here, f denotes the original model being explained, g represents the explanation model, G is the set of candidate explanation models, L is a loss function measuring local disagreement between the two models, π_x represents the weights assigned to the neighborhood of the instance being explained, and $\Omega(g)$ represents the complexity of the explanation model. The objective is to obtain a model that adequately represents the original model's behavior around the instance of interest while remaining interpretable.

A major advantage of LIME is that it can be applied to different classifiers without being restricted to a particular model family. In text classification, local explanations can identify words that support or weaken a prediction. In image classification, explanations can be generated using specific image regions.

However, LIME's results may vary depending on the sampling strategy, how the local neighborhood is defined, and the complexity of the explanation model. Moreover, the fact that a local explanation model performs well around a particular instance does not mean that it accurately represents the original model's behavior throughout the entire input space. LIME explanations should therefore be evaluated for their stability under different settings and their agreement with the original model (Ribeiro et al., 2016; Carvalho et al., 2019; Meng et al., 2024).

3.3. Feature Importance Analysis

Feature importance analysis encompasses methods used to investigate which variables play a greater role in model predictions. Such analyses can help identify the inputs on which a model relies, detect potential data problems, and investigate model behavior (Carvalho et al., 2019). One widely used approach is permutation-based feature importance, which examines how much model performance changes after the values of a variable are shuffled. If performance declines after shuffling, this may indicate that the model relies on the corresponding variable when making predictions (Molnar, 2022).

However, importance scores do not directly indicate the causal effects of variables in the real world. Correlated features may contain similar information within a model, making their importance scores more difficult to interpret. Furthermore, the fact that a variable is important to a model does not establish that decisions based on that variable are fair or accurate (Ebrahimi et al., 2025). Rather than using feature importance analysis as a standalone decision criterion, it is therefore more appropriate to treat it as a starting point for understanding overall model behavior. In applications involving consequential decisions, such analyses should be supplemented with instance-level investigations and independent performance tests.

3.4. Counterfactual Explanations

Counterfactual explanations aim to identify the changes to input features that might be required to produce an outcome different from the current prediction. While feature attribution methods address the question, “Which features are associated with this prediction?”, counterfactual methods focus on the question, “Which changes could be associated with a different outcome?” (Wachter et al., 2018; Guidotti, 2024).

A basic counterfactual explanation problem can be formulated as shown in Equation 3:

$$x' = \underset{z}{\operatorname{argmin}} d(x, z) \quad \text{subject to} \quad f(z) = y_{\text{target}} \quad (3)$$

Here, x represents the current instance, z denotes an alternative instance, $d(x, z)$ measures the distance between the two instances, f is the prediction model, and y_{target} is the desired outcome. The objective is to find an alternative instance that produces the target output while remaining as close as possible to the original instance.

Counterfactual explanations can be particularly useful when users need actionable information. However, a change that appears mathematically feasible may not be practical in real life. Explanations that recommend changing immutable features, such as age or events that have already occurred, are also unsuitable in practical terms. Changeability, realism, and relevance to the user should therefore be considered when evaluating counterfactual explanations (Guidotti, 2024).

It is also important not to confuse counterfactual explanations with causal relationships in the real world. The fact that a model produces a different output for a different input does not, by itself, prove that making the same change in real life would produce the same outcome. These explanations should therefore primarily be regarded as tools for examining model behavior.

3.5. Gradient-Based and Visual Explanation Methods

Some explanation methods used with DL models exploit the sensitivity of model outputs to changes in their inputs. Gradient-based methods examine the relationship between changes in the output and components of the input to identify the features to which the model is sensitive (Wang et al., 2024). Integrated Gradients evaluates changes in a model's output along a path between an input

and a reference input by using gradients. The method provides an axiomatic approach to addressing situations in which ordinary gradients may be insufficient (Sundararajan et al., 2017). Gradient-weighted Class Activation Mapping (Grad-CAM), in turn, aims to display image regions associated with a particular class prediction in CNNs as a heatmap. This method makes it possible to visually examine which regions of an image may contribute to a classification result (Selvaraju et al., 2017).

These visualizations can make it easier to investigate the image regions to which a model is sensitive. However, highlighting a particular region in a heatmap does not prove that the model understands that region in the same way humans do or that its decision necessarily depends on that region. Whether the explanation reflects the model's learned behavior must be tested separately (Adebayo et al., 2018). The reliability of gradient-based and visual explanation methods should therefore be assessed not merely by inspecting the visualizations they produce, but by systematically testing whether the explanations are consistent with the behavior learned by the model.

3.6. Attention Mechanisms and Relationship to Explanation

Attention mechanisms, particularly in natural language processing and transformer architectures, involve computations that indicate how a model distributes weights across different input components. These weights can be used to examine which components receive greater weight during processing. However, interpreting attention weights directly as explanations of model decisions remains controversial (Jain & Wallace, 2019; Lyu et al., 2024).

Jain and Wallace (2019) presented evidence showing that attention weights cannot always be regarded as reliable explanations for model predictions. A high attention weight assigned to a word or

text segment does not, by itself, establish its influence on the final prediction.

Attention weights should therefore be treated as indicators that can help investigate a model's internal computations. Their validity as explanations must be independently established for the particular model and task.

3.7. Rule-Based and Example-Based Explanations

Rule-based explanations aim to describe the conditions under which a decision is made in an understandable form. Example-based methods, on the other hand, help investigate model decisions by relating a prediction to similar instances. These approaches can be useful, particularly when explanations need to be understood by non-technical users (Guidotti et al., 2018; Hong & Park, 2025).

The Anchors approach developed by Ribeiro et al. (2018) focuses on identifying sets of high-precision conditions under which a particular prediction remains valid. However, a rule that achieves high precision for certain instances is not necessarily valid across a broader set of examples. Similarly, two instances that a model considers mathematically similar may not be meaningfully similar from a human perspective.

Accordingly, when using rule-based and example-based explanations, it is important to consider the conditions under which an explanation remains valid, its scope, and the suitability of the similarity measure being used.

4. Comparison of Explanation Methods

Because explanation methods address different questions, it would be inappropriate to claim that a single method is the best choice for every application. SHAP and LIME help examine feature contributions, whereas counterfactual explanations focus on the changes required to obtain alternative outcomes. Methods such as

Integrated Gradients and Grad-CAM can be used to investigate the relationship between inputs and the outputs of DL models. The main characteristics of these explanation methods are summarized comparatively in Table 1 (Ribeiro et al., 2016; Lundberg & Lee, 2017; Selvaraju et al., 2017; Sundararajan et al., 2017; Adebayo et al., 2018; Guidotti et al., 2018; Ribeiro et al., 2018; Jain & Wallace, 2019; Molnar, 2022; Guidotti, 2024).

Table 1 Comparison of the main characteristics of explanation methods

Method	Main Purpose	Advantage	Limitation
SHAP	Attribute feature contributions to predictions	Can be used for individual and aggregate analyses	Depends on assumptions and computational cost
LIME	Explain a specific prediction using a local model	Applicable to different model types	Results may vary with sampling and parameter settings
Feature importance analysis	Examine variables considered important by the model	Provides insights into overall behavior	Does not establish causal effects
Counterfactual explanations	Identify changes associated with an alternative outcome	Can provide actionable information	Suggested changes may be unrealistic or impractical
Integrated Gradients	Examine input contributions using gradients	Can be used with deep neural networks	May be sensitive to the reference input
Grad-CAM	Visualize regions associated with an image prediction	Facilitates visual inspection	Does not, by itself, establish the reason for a decision
Attention weights	Examine the distribution of attention	Provides an indication of internal computations	Not always a reliable explanation
Rule-based and example-based methods	Explain decisions through conditions or examples	Can be understandable to humans	May involve limitations in scope and similarity

In practice, method selection should begin by identifying the question that needs to be explained. Feature importance analysis may be appropriate when the objective is to investigate which variables the model generally relies on. SHAP or LIME may be considered when the rationale behind a specific prediction is of interest. Counterfactual explanations may be more suitable when the aim is to inform users about how an alternative outcome might be achieved.

It is also possible to combine multiple methods. For example, general feature importance can be examined first, followed by SHAP explanations for selected instances. Nevertheless, different methods should not be expected to produce identical results, as they rely on different assumptions and explanatory objectives. Differences between explanations can therefore serve as an indication that the model and the methods themselves warrant further investigation.

5. Application Domains and Current Challenges

5.1. Healthcare and Clinical Decision Support

AI can be used in healthcare for tasks such as medical image analysis, disease risk prediction, and clinical decision support. In these applications, it is important not only for a model to produce accurate predictions but also for users to understand how those predictions should be assessed in a clinical context. Explanation methods can help healthcare professionals examine model outputs and identify potential errors (Arrieta et al., 2020; Saw et al., 2025).

For example, the image regions to which an image classification model is sensitive can be examined using methods such as Grad-CAM. However, the fact that a visualization highlights a clinically meaningful region does not, by itself, prove that the model identified the disease for the correct reasons. The model may also rely on image characteristics that are not directly related to the disease. Explanations should therefore not replace clinical validation or independent performance evaluations (Selvaraju et al., 2017; Adebayo et al., 2018).

The intended audience is also important in this domain. The technical details required by a researcher may differ from the information a clinician needs when making a decision. Explanations intended for patients may require a different language and level of detail. The target user and decision-making context should therefore be taken into account when designing explanations.

5.2. Finance and Credit Assessment Systems

In the financial sector, ML can be used for tasks such as credit risk prediction, fraud detection, and risk management. In these applications, explanations can support the examination of features associated with a particular prediction and the investigation of potential model errors. For example, SHAP or LIME can be used to examine the contributions of specific variables to a credit application assessment (Ribeiro et al., 2016; Lundberg & Lee, 2017).

However, the importance of a variable in a prediction does not establish that a decision is fair or legally compliant. If the training data reflect historical inequalities, explanation methods may help investigate these patterns, but they do not eliminate them automatically. Fairness assessments must also consider model performance across different groups, data quality, and compliance with the rules governing the decision-making process.

Counterfactual explanations can also be used to provide information to users in financial applications. For example, they can indicate the conditions under which a model might produce a different decision. Nevertheless, the suggested changes must be realistic and feasible. A mathematical scenario that changes a model's output may not translate into a recommendation that a user can implement in real life (Wachter et al., 2018; Guidotti, 2024).

5.3. Cybersecurity and Anomaly Detection

In cybersecurity, ML is used for tasks such as network traffic classification, malware detection, and the identification of unusual behavior (Mohale & Obagbuwa, 2025). Explanation methods can help investigate why an event was classified as suspicious or which features increased its risk score (Guidotti et al., 2018; Calzarossa et al., 2025). Security analysts can use these explanations to compare model outputs with independent indicators and investigate potential false positives.

However, explainability has a particular dimension in cybersecurity: detailed information about a model's decision logic can support defensive operations, but it can also help malicious actors understand the system's limitations. It is therefore important to consider who receives an explanation and the level of detail shared. Explainability should form part of security assessment rather than replace robustness testing, error analysis, or operational monitoring.

5.4. Public Services and High-Stakes Decisions

Automated decision systems used in public services, education, employment, and social assistance can affect individuals' rights and opportunities. Understanding the reasons behind decisions, investigating errors, and auditing decision-making processes are therefore important. In high-stakes applications, implementing a system that generates post-hoc explanations may not be sufficient; the use of directly interpretable models should also be considered whenever feasible (Rudin, 2019).

The European Union's Artificial Intelligence Act establishes a regulatory framework that includes risk-based obligations and transparency-related requirements for certain AI systems. However, the regulation should not be interpreted as imposing the same unrestricted explanation requirements on every AI system. The applicable requirements depend on the system's classification, intended purpose, and relevant provisions (European Parliament & Council of the European Union, 2024).

The Artificial Intelligence Risk Management Framework published by the US National Institute of Standards and Technology (NIST) also links explainability and interpretability to the management of risks associated with trustworthy AI. This approach supports the view that explainability should not be reduced to a single technical method but should instead be incorporated into the

design, development, evaluation, and use of AI systems (NIST, 2023).

5.5. Explainability, Trust, and User Evaluation

One of the major challenges in XAI is the assumption that providing more explanations will always increase trust. In practice, however, the goal should not simply be to increase trust but to ensure that it is appropriately calibrated. Excessive trust in an unreliable model, as well as the unjustified rejection of a reliable one, can lead to undesirable outcomes (Miller, 2019; NIST, 2023).

The benefits explanations provide to users should not be measured solely by whether users like them or find them easy to understand. It is also necessary to examine whether explanations improve decision quality, help users understand model limitations, and prevent misplaced confidence. The literature on XAI evaluation emphasizes that user-centered evaluation constitutes a distinct dimension that complements technical criteria (Miller, 2019; Naveed et al., 2024).

Accordingly, two fundamental questions should be considered together when evaluating explanations: Does the explanation accurately reflect the model's behavior? Can the intended user understand and use the explanation appropriately? Answering only one of these questions is insufficient to assess the overall quality of an explanation.

5.6. Current Challenges and Research Directions

One of the major challenges in XAI is establishing context-appropriate criteria for evaluating the quality of different explanation methods. Because these methods have different objectives, comparing all of them using a single metric is difficult. For example, evaluating a feature attribution method and a method that identifies the changes required to obtain alternative outcomes using the same

criterion may overlook their fundamental differences (Ding et al., 2022; Schwalbe & Finzel, 2024).

Another challenge concerns the stability and reliability of explanations. Producing substantially different explanations for similar inputs can make it difficult for users to understand model behavior. Nevertheless, consistency alone is not sufficient; an explanation must also be sensitive to the model's actual behavior. Explanation methods should therefore be tested both for stability and for their ability to represent model behavior accurately (Adebayo et al., 2018; Carvalho et al., 2019; Calzarossa et al., 2025; Fridkin & Bendersky, 2026).

The relationship between explainability and causality also requires careful consideration. Feature contributions, gradients, and changes in model predictions can provide information about how a model behaves, but they do not independently establish real-world cause-and-effect relationships. Explaining model behavior must therefore be distinguished from identifying causal mechanisms in the real world (Wachter et al., 2018; Baron, 2023).

The growing adoption of generative AI and large language models (LLMs) has introduced new questions for explainability research. The complex internal structures of these systems can make it difficult to fully explain the processes through which a particular response is generated. Although attention weights and feature attribution methods can be useful for specific investigations, they should not be assumed to provide a complete explanation of the model's internal workings or the correctness of its responses (Jain & Wallace, 2019; Fridkin & Bendersky, 2026).

Finally, explainability should not be addressed only at the final stage of model development. Different explanatory needs may arise during data preparation, model development, testing, deployment, and post-deployment monitoring. The selection of

explanation methods should therefore be planned alongside model documentation, risk management, and user evaluation (NIST, 2023).

6. Conclusion and Future Research

XAI is an important research field that aims to make the predictions of complex ML models more understandable to humans. Feature attribution methods such as SHAP and LIME help investigate individual predictions, while feature importance analyses provide insights into overall model behavior. Counterfactual explanations identify changes that may lead to alternative outcomes, whereas gradient-based and visual methods can be used to investigate the relationship between inputs and DL model outputs.

Each of these methods has its own assumptions, conditions of use, and limitations. An explanation method should therefore not be selected solely because it is widely used or produces visually understandable outputs. The question to be explained must first be identified, the suitability of the method must be tested, and the results must be evaluated in the context in which they will be used.

An understandable explanation does not establish that a model is accurate or fair. Explanation methods can account for incorrect decisions and may, in some cases, represent model behavior only approximately. The technical reliability of explanations and the benefits they provide to users should therefore be evaluated separately.

Future research will continue to focus on evaluating explanation methods more consistently, developing forms of explanation that meet user needs, and clarifying the boundaries between causal explanations and explanations of model behavior. In addition, explainability should be considered alongside fairness, robustness, accountability, and risk management.

In conclusion, XAI provides tools that contribute to understanding black-box models, but it is not sufficient on its own to ensure trustworthy AI. An effective approach requires selecting an appropriate explanation method, testing the resulting explanations, and evaluating the findings in the relevant decision-making context. Particularly in high-stakes applications, explainability should be treated as part of a broader assessment and governance process that extends throughout the entire model lifecycle.

References

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, pp. 9505–9515).

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>

Baron, S. (2023). Explainable AI and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33, 347–377. <https://doi.org/10.1007/s11023-023-09637-x>

Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review*, 58, Article 150. <https://doi.org/10.1007/s10462-025-11141-w>

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>

Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I., & Sannino, G. (2025). Ante-hoc methods for interpretable deep models: A survey. *ACM Computing Surveys*, 57(10), Article 262, 1–36. <https://doi.org/10.1145/3728637>

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>

Ebrahimi, S., Abdelhalim, E., Hassanein, K., & Head, M. (2025). Reducing the incidence of biased algorithmic decisions through feature importance transparency: An empirical study.

European Journal of Information Systems, 34(4), 636–664.
<https://doi.org/10.1080/0960085X.2024.2395531>

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

Fayyaz, A. M., Abdulkadir, S. J., Talpur, N., Al-Selwi, S. M., Hassan, S. U., & Sumiea, E. H. (2025). Grad-CAM (Gradient-weighted Class Activation Mapping): A systematic literature review. *Computers in Biology and Medicine*, 198, Article 111200. <https://doi.org/10.1016/j.compbiomed.2025.111200>

Fridkin, S., & Bendersky, M. (2026). Interpretable machine learning: A comprehensive review of foundations, methods, and the path forward. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 16(1), Article e70075. <https://doi.org/10.1002/widm.70075>

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>

Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38, 2770–2824. <https://doi.org/10.1007/s10618-022-00831-6>

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>

Hong, S., & Park, W. (2025). Developing user-centered system design guidelines for explainable AI: A systematic literature review. *Artificial Intelligence Review*, 58, Article 386. <https://doi.org/10.1007/s10462-025-11363-y>

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3543–3556). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1357>

Lin, T., Lau, R. Y. K., & Hu, S. (2026). A critical review of state-of-the-art explainable artificial intelligence (XAI) methods and their business applications. *Artificial Intelligence Review*, 59, Article 161. <https://doi.org/10.1007/s10462-026-11565-y>

Lipton, Z. C. (2018). The mythos of model interpretability. *ACM Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>

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).

Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2), 657–723. https://doi.org/10.1162/coli_a_00511

Meng, H., Wagner, C., & Triguero, I. (2024). SEGAL time series classification—Stable explanations using a generative model and an adaptive weighting method for LIME. *Neural Networks*, 176, Article 106345. <https://doi.org/10.1016/j.neunet.2024.106345>

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>

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, 8, Article 1526221. <https://doi.org/10.3389/frai.2025.1526221>

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>

Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences*, 14(23), Article 11288. <https://doi.org/10.3390/app142311288>

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>

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>

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>

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), Article 2400304. <https://doi.org/10.1002/aisy.202400304>

Saw, S. N., Yan, Y. Y., & Ng, K. H. (2025). Current status and future directions of explainable artificial intelligence in medical imaging. *European Journal of Radiology*, 183, Article 111884. <https://doi.org/10.1016/j.ejrad.2024.111884>

Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38, 3043–3101. <https://doi.org/10.1007/s10618-022-00867-8>

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). <https://doi.org/10.1109/ICCV.2017.74>

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 3319–3328). PMLR.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.

Wang, Y., Zhang, T., Guo, X., & Shen, Z. (2024). Gradient based feature attribution in explainable AI: A technical review. *arXiv*. <https://doi.org/10.48550/arXiv.2403.10415>