Explainable Artificial Intelligence (XAI): Bridging the Gap Between Accuracy and Interpretability

Dr. Anika P. Moreau
Institute for Trustworthy Machine Intelligence, Université Paris-Saclay, Paris, France
Received: 27 /01/ 2026 ; Accepted: 22 /03/2026 ; Published: 23/04/2026

Abstract
The increasing demand for trust, accountability, and transparency in sophisticated machine learning systems has led to the rise of Explainable AI (XAI) as a significant field of study. Their "black-box" aspect frequently limits interpretability and hinders their adoption in high-stakes decision-making situations, even if advanced models, especially deep learning architectures, have accomplished outstanding accuracy across domains including healthcare, autonomous systems, and finance. It is a major issue for researchers and practitioners alike to strike a balance between model transparency and prediction performance. To close this gap, XAI is working on ways to explain model behavior to people in a way that does not sacrifice accuracy. To shed light on the prediction generation process, researchers have turned to techniques like feature importance analysis, attention processes, model-agnostic explanation methods (like LIME and SHAP), and intrinsically interpretable models. Model judgments can be validated, biases can be detected, and regulatory and ethical compliance can be assured with the use of these methodologies.
Keywords Explainable Artificial Intelligence (XAI) , Interpretability , Model Transparency , Black-Box Models

Introduction
A number of AI applications, such as healthcare diagnostics, financial forecasts, autonomous vehicles, and natural language processing, have seen state-of-the-art performance from machine learning and deep learning models in recent years. Highly accurate predictions can be made by these models, especially those with complicated structures like deep neural networks, which can capture detailed patterns in massive datasets. The "black-box" problem in artificial intelligence arises when, despite the greater accuracy, interpretability suffers as a result. Decisions in high-stakes domains need to be explicable, dependable, and responsible, but black-box models can not provide that. For example, AI-driven diagnoses in healthcare must have clear clinical reasoning, and automated decision-making systems in finance must be transparent due to legal requirements. Trusting, validating, or auditing the results produced by these models becomes challenging in the absence of sufficient explanations. This has sparked rising worries about the ethical implications of AI deployment, as well as issues of justice and bias. The rise of Explainable AI (XAI) is a reaction to these difficulties; it seeks to improve the interpretability and user-friendliness of machine learning models. In order to shed light on the process by which models generate their predictions, XAI incorporates a wide range of methodologies and procedures. Among these, you can find post hoc explanation approaches like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), and intrinsically interpretable models that prioritize transparency in their construction. Closing the gap between accurate models and interpretable results is XAI's

principal goal. Contemporary XAI methods aim to keep high accuracy while improving transparency and user confidence, instead than compromising predictive performance for simplicity. Maintaining this equilibrium is critical for the ethical and legal implementation of AI systems, especially in contexts where rules like the General Data Protection Regulation (GDPR) stress the "right to explanation." When it comes to detecting and reducing biases in machine learning models, XAI is just as important as making things more transparent. Explainability techniques help stakeholders find biased or unfair patterns in data or algorithms by revealing the relative relevance of features and decision paths. This helps advance AI systems that are more equitable and welcoming to everybody. Explore the concepts, methods, and uses of Explainable AI, with an emphasis on its solutions to the problem of balancing precision with accessibility. It takes stock of current approaches, assesses how well they work, and calls attention to present difficulties and potential avenues for further study. The research aspires to aid in the development of AI systems that are trustworthy, transparent, and morally aligned through this approach.

Concept and Importance of Explainable Artificial Intelligence (XAI)

What we call "Explainable AI" (XAI) is really just a collection of practices aimed at making AI systems' decisions and outputs more human-friendly. The inner workings of machine learning models, especially deep learning systems, are frequently opaque, making it difficult to understand how given inputs result in particular outputs, especially as these models grow in complexity. XAI solves this problem by making model behavior transparent, so people can understand the logic behind AI-driven decisions.

Essential to XAI is the goal of making "black-box" models more interpretable through the creation of models with inherent transparency or the application of post hoc explanation techniques to complicated models. To help consumers understand which factors impacted a certain prediction, these explanations may provide feature importance scores, decision rules, visualizations, or local approximations. We want to make sure that the explanations we provide for outcomes are meaningful, accurate, and helpful for decision-making, not just that we explain them.

The broad use of AI across important areas has greatly increased the significance of XAI. For example, in healthcare, doctors need to know why AI-assisted diagnoses are correct so they can give patients the best care possible. Compliance with regulations and risk assessment in the financial sector necessitate that institutions provide explanations for automated decisions like loan approvals and credit scoring. The same holds true in governance and legal contexts: openness is key to holding algorithms accountable and making their decisions fairly.

Building trust between people and AI systems is one of XAI's primary accomplishments. Users are more inclined to trust a model's results when they can comprehend how the model reaches its conclusions. In situations where biased or wrong decisions could have major repercussions, this is of the utmost importance. Users are able to test model predictions, see any mistakes, and make educated decisions on when to trust or overrule AI recommendations with the support of XAI's clear explanations.

A further essential feature of XAI is its capacity to identify and reduce bias in ML models. Unfair or discriminating results may be produced by AI systems that have been trained using biased or inaccurate data. To better detect and eliminate these biases, researchers and

practitioners can use explainability techniques to investigate how various features impact model predictions. This helps advance AI systems that are more fair and accountable.

When it comes to following the rules, both ethical and legal, XAI is an absolute must. The requirement for openness and the "right to explanation" in automated decision-making systems is emphasized by regulations such as the General Data Protection Regulation (GDPR). So, businesses using AI need to make sure their models are accurate, interpretable, and accountable.

In addition, XAI optimizes model improvement and debugging. Gaining insight into a model's decision-making process allows developers to pinpoint areas for improvement, enhance the model's architecture, and boost performance. In addition to connecting complicated algorithms to real-world applications, it improves communication between technical specialists and stakeholders without technical backgrounds.

Need for Explainability in AI Systems

The requirement for AI systems to be explainable has grown in importance because to their increasing integration into decision-making processes in industries including healthcare, finance, governance, and transportation. Although deep learning systems and other advanced machine learning models provide very accurate predictions, it is not always easy to decipher their decision-making processes due to their complexity and opaqueness. The immediate necessity for explainable AI (XAI) is brought to light by the fact that this opaqueness poses problems with trust, responsibility, and moral implementation.

Building trust and confidence among users is a key motivation for make things easy to understand. When the logic behind AI systems' forecasts and recommendations is clear, stakeholders have more faith in them. For instance, a doctor is not going to take an AI system's diagnosis at face value until the system explains all the elements that went into making that determination. Users' ability to validate results and make educated judgments is greatly enhanced by explainability, which in turn strengthens confidence in AI systems.

Responsibility and accountability are further critical factors. The results of AI system decisions can be disastrous in situations where the stakes are very high. When mistakes happen, it is crucial to find out what went wrong. Developers, companies, and regulators are able to review decisions and assign responsibilities correctly with explainable models. The lack of explainability makes it hard to tell if the problem is with the data, the model, or the way it was implemented.

The ability to understand and explain something is also critical for fighting prejudice and bigotry. Unfair or discriminatory results could be produced by AI systems trained on biased datasets, especially in sectors like recruiting, lending, and law enforcement. More fair AI systems can be achieved with the use of explainability techniques, which shed light on feature importance and decision processes, allowing for the detection of bias tendencies and the encouragement of remedial actions.

Furthermore, openness in automated decision-making is greatly emphasized by regulatory and legal obligations. People have an entitlement to know the reasoning behind actions that impact them, according to frameworks like the General Data Protection Regulation (GDPR). In order to stay in compliance with these requirements, organizations using AI must make sure their systems explain things clearly and in a way that anyone can understand.

Model validation and improvement also heavily rely on explainability. Researchers and developers employ interpretability techniques to fix bugs, find mistakes, and enhance algorithms. Improving performance, eliminating overfitting, and guaranteeing robustness across multiple datasets and circumstances can be achieved by understanding how a model arrives at its predictions.

In addition, human-AI collaboration is improved by explainability. Artificial intelligence systems are often developed to supplement human decision-makers, not to supplant them. By providing detailed explanations, humans are able to better understand AI outputs, integrate them with their domain knowledge, and ultimately make more educated decisions. To get the most out of AI while keeping the downsides to a minimum, this team effort is crucial.

Lastly, explainability encourages openness, justice, and responsibility, which are all necessary for the creation of ethical AI. The public's trust and the prevention of misuse depend on AI systems operating in a transparent and explicable manner, since these systems are having an ever-increasing impact on society outcomes.

Types of Explainability: Global vs Local Interpretability

There are two main varieties of XAI that are complementary to one another: global interpretability and local interpretability. Despite their differences in emphasis and methodology, when combined, these methods illuminate the inner workings of machine learning models and their predictive abilities in great detail.

Global Interpretability

To be globally interpretable, a machine learning model must be able to have its general logic and behavior understood throughout the whole dataset. Some of the questions it hopes to address include: What is the model's decision-making process like in general? therefore, in general, which characteristics have the greatest impact?

Models like rule-based systems, decision trees, and linear regression are simpler and inherently interpretable, therefore they tend to exhibit this kind of explainability. It is possible to see and study the connection between input properties and outcomes in these models. Approximation methods, feature importance rankings, or surrogate models that imitate the original system's behavior can accomplish global interpretability for increasingly complicated models.

Key characteristics of global interpretability:

- Provides a holistic view of model behavior
- Identifies overall feature importance and trends
- Useful for model validation, debugging, and policy-level decisions
- Helps ensure consistency and fairness across the dataset

However, global explanations may oversimplify complex models and fail to capture nuanced behaviors in specific cases.

Local Interpretability

The goal of local interpretability is to provide context for specific model predictions. Questions like "why did the model make this specific decision for a particular input?" are addressed rather than the model's overall behavior.

Given the difficulty in gaining a global knowledge of large "black-box" models like deep neural networks, this method is particularly crucial for these types of models. Local interpretable

model-agnostic explanations (LIME) and SHAP (SHapley Additive exPlanations) are two popular methods for this purpose. They estimate the model's behavior around a given data point and produce local explanations.

Key characteristics of local interpretability:

- Describes specific choices or forecasts
- Emphasizes the contributions to features for particular cases
- Perfect for situations when the outcome is crucial, such as in the medical and financial fields.
- Improves user-level trust and accountability

Despite its utility, local interpretability may not provide insights into the model's overall behavior and can occasionally produce inconsistent interpretations across different cases.

Comparison and Complementarity

For a comprehensive grasp of AI systems, both global and local interpretability are crucial. At the system level, global interpretability guarantees transparency, whereas at the individual decision level, local interpretability gives precise insights. By integrating the two methods, stakeholders can assess the efficacy of models, guarantee equity, and establish confidence in AI systems from a more well-rounded viewpoint.

It is typically inadequate to depend on a single form of explanation in contemporary AI systems, particularly those built on deep learning. Better, more open, and accountable machine learning models are possible using a hybrid approach that combines global and local interpretability; this, in turn, helps with the ethical use of AI in the real world.

Techniques in Explainable AI

To improve the interpretability and transparency of machine learning models, Explainable AI (XAI) uses a number of methods. There are two main ways to classify these methods: those that are applied to models that are intrinsically interpretable and those that are used as post hoc procedures to explain sophisticated "black-box" systems. Depending on the kind of model and the context of application, each technique provides varying degrees of insight.

1. Feature Importance Methods

Finding out which input variables have the biggest impact on a model's predictions is the goal of feature importance techniques. These techniques rate features according to how much of an impact they have on the final product, either on a global or local scale.

- **Global feature importance** evaluates the overall contribution of each feature across the dataset.
- **Local feature importance** explains the role of features in a specific prediction.

Permutation significance and attribution methods based on gradients are two common approaches. Validation of models, debugging, and bias identification are common applications of these techniques.

2. Model-Agnostic Explanation Methods

No matter the internal structure of a machine learning model, model-agnostic approaches can be used with it.

- **LIME (Local Interpretable Model-Agnostic Explanations):** Approximates a complex model locally using a simpler interpretable model to explain individual predictions.
- **SHAP (SHapley Additive exPlanations):** To guarantee consistent and theoretically sound explanations, SHAP uses game theory to give contribution values to each feature.

Predictions made by deep learning and ensemble models can be better understood with the help of these techniques.

3. Inherently Interpretable Models

The decision-making process of certain models is straightforward because of their inherent transparency.

- A comparison of logistic and linear regression
- Recursive models and decision trees
- The GAMs are all-purpose additive models.

Although these models make the inputs and outputs straightforward, they may not be as accurate when dealing with complicated data.

4. Visualization Techniques

Visualization methods help interpret model behavior through graphical representations.

- **Partial Dependence Plots (PDPs):** Show how a feature affects predictions on average.
- **Individual Conditional Expectation (ICE) plots:** Display how predictions vary for individual data points.
- **Saliency maps and heatmaps:** Highlight important regions in images for computer vision models.

Particularly for stakeholders without technical knowledge, these methods simplify complicated model behaviors.

5. Attention Mechanisms

Deep learning models rely heavily on attention mechanisms, especially for tasks like computer vision and natural language processing. They draw attention to the specific pieces of input data that the model employs for decision-making.

Attention scores, for instance, show which words are most important for next-word prediction and context interpretation in language models. As a result, there is some inherent interpretability.

6. Surrogate Models

Trained to mimic the behavior of more complicated models, surrogate models are easier to understand and work with.

- By simulating the original model's overall behavior, global surrogate models are created.
- Forecasts in a limited area of the input space can be better understood with the help of local surrogate models.

While they do a good job of shedding light on complicated systems, they could miss some subtleties in the original model.

7. Rule-Based and Example-Based Explanations

- **Rule-based explanations** deduce principles that can be understood by humans from models, therefore facilitating the interpretation of decisions.

Example-based explanations utilize comparable historical examples to substantiate forecasts, facilitating user comprehension of results through comparison.
Domains that rely heavily on human reasoning and justification are ideal for these methods.
Model complexity, application domain, and desired interpretability level dictate the XAI technique to be used. A more complete knowledge of AI systems is typically achieved by integrating numerous methodologies, while no single method is adequate in all cases. Research toward explainability methods that are more user-centric, scalable, and robust is important because AI is always changing.

Conclusion

With the growing influence of AI systems on important decisions in many different fields, Explainable Artificial Intelligence (XAI) has emerged as a crucial component in AI system development and deployment. The lack of openness surrounding powerful machine learning models has sparked major issues about trust, accountability, fairness, and ethical use, despite their impressive levels of accuracy. To solve these issues, XAI has developed methods to simplify and simplify complicated models so that they may be understood and used by humans. Stakeholders can learn more about the decision-making process with the help of XAI's many techniques, such as feature importance analysis, visualization tools, model-agnostic procedures, and inherently interpretable models. Model validation, bias detection, and regulatory compliance are all helped along by this, and it also boosts trust and user confidence. The explainability framework is enhanced by the differentiation between global and local interpretability, which provides insights at both the system and instance levels. In spite of all the progress, XAI still has a ways to go before it can solve all of its problems. Some of these problems include making sure explanations are accurate and relevant to all users, and striking a balance between interpretability and predictive performance. Furthermore, more scalable and strong explainability methods are required due to the rising complexity of AI models. Future work in XAI should focus on hybrid techniques that build user-centric explanation systems for various stakeholders and on incorporating transparency into high-performing models directly. Explainability is going to be crucial in making sure that AI systems are accurate, fair, transparent, and accountable as legislative frameworks change and more people want AI to be used responsibly.

References (APA Style)

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

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

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

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), 93. https://doi.org/10.1145/3236009

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

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*.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Lulu.com.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *Berkman Klein Center for Internet & Society Research Paper*.