Explainable Artificial Intelligence

Dominika Darabos

October 14, 2025

Introduction

The remarkable performance of deep neural networks has driven their adoption across a wide range of domains. At the same time, the black-box nature of their internal processes has raised significant concerns regarding transparency, trust, and accountability. Explainable artificial intelligence (XAI) has emerged in response, aiming to "open up the black box" by providing insights into how models generate their decisions. In practice, XAI encompasses methods and processes that allow humans to understand, interpret, and ultimately trust the outputs of machine learning systems. Explanations are particularly vital in safety-critical areas such as healthcare, transportation, finance, law, and military systems [1], where incorrect predictions can have severe consequences. By making model reasoning more transparent, erroneous predictions can be detected and addressed, leading to safer and more responsible AI usage. Moreover, explanations can highlight structural weaknesses in models, creating opportunities for improvement. Historically, safety-critical domains have relied on simpler but more transparent models, even at the cost of reduced accuracy. XAI offers a way to reconcile this trade-off, bridging the gap between the high performance of modern neural networks and the need for interpretability in critical applications.

As Lipton [2] notes, interpretability methods in AI fall broadly into two categories. Transparency aims to reveal the full internal logic of a model, enabling decisions to be reproduced without the model itself. While this represents the strictest form of interpretability, it is often impractical for modern, highly complex systems and not always necessary in practice.

By contrast, post-hoc interpretability focuses on clarifying the key factors behind a decision without exposing every computational step. Explanations may take the form of natural language, visualizations, or case-based reasoning, selectively highlighting the most influential factors that contributed to an outcome. This human-centered selectivity mirrors how people naturally explain decisions, making post-hoc methods both practical and widely applicable [3].

Building on these foundations, Adadi and Berrada [1] distinguish between model-specific and model-agnostic interpretability. Model-specific methods are constructed especially for a particular architecture, while model-agnostic approaches can be applied to any model, making them valuable for cross-model

comparison. A second distinction concerns the scope of explanation: global methods aim to clarify the overall decision logic of a model, whereas local methods explain individual predictions.

Among post-hoc, model-agnostic approaches, Local Interpretable Model-agnostic Explanations (LIME) explains individual predictions by approximating the model's behavior in the neighborhood of a given instance with a simple, interpretable model, such as a sparse linear regression [4]. It generates slightly modified versions of the input, observes how the predictions change, and then weights these samples based on their closeness to the original point. In this way, LIME highlights which features most influenced the decision, offering practical insights without requiring access to the model's internal structure.

While model-agnostic techniques such as LIME provide general insights, other approaches directly exploit the structure of deep neural networks. Two of the most prominent are Layer-Wise Relevance Propagation (LRP) and Integrated Gradients (IG), both of which produce relevance scores indicating how much each part of the input contributes to a prediction.

LRP uses backward propagation as each neuron assigns its relevance to earlier neurons in proportion to their forward contribution [5], until all relevance is assigned to the input features.

IG [6] is based on two axioms: sensitivity, which ensures that features affecting the prediction receive non-zero attribution, and implementation invariance, which guarantees that two models computing the same function yield identical explanations. To satisfy these principles, IG computes attributions by integrating gradients of the model's output with respect to the inputs along a straight path from a baseline (e.g., a zero vector) to the actual input. The result is a set of relevance scores that reflect how the prediction changes relative to the baseline.

DeepLIFT assigns attributions by comparing neuron activations to those from a reference input and propagating the differences backward through the network, allowing it to capture contributions even when gradients vanish [7]. The SHAP framework further advanced the field by linking attribution methods to Shapley values from cooperative game theory [8]. By enforcing properties such as local accuracy, missingness, and consistency, SHAP established a principled definition of feature attributions and showed that widely used techniques like LIME, IG, and DeepLIFT can be understood as approximations of Shapley values.

Taken together, these methods demonstrate the range of approaches for explaining complex models. Each produces relevance scores that quantify the contribution of input features to predictions, providing insight into model behavior.