AI Explainability Techniques 101: Methods for Interpreting AI Models

TL;DR

A variety of techniques can help shed light on the complex inner workings of AI systems. By increasing the transparency of AI decision-making, you can gain better control over AI systems while increasing user trust and regulatory compliance across the enterprise.

Introduction

Modern AI systems do extraordinary things, but it can be difficult to understand how or why they do them. This is why AI systems are often labelled "black boxes" — because it is difficult to understand or explain their inner workings. Explainable AI (XAI) is the project of finding techniques and practices to address these challenges and open up the AI black box.

Successful AI explainability techniques increase the transparency and interpretability of AI systems. This is vital for a variety of reasons. Interpretable AI systems are easier to control for both users and developers. It can be difficult to fine-tune the behaviour of a model without a grasp of how and why the model does what it does. Transparent AI systems also increase user and developer trust in the decisions and outputs of AI systems.

This guide explains several important AI explainability techniques and everything you need to know to get started in this developing area of AI technology.

Key Takeaways

1. AI explainability techniques enable developer control, increase user trust and facilitate compliance with AI regulation.

2. Some AI algorithms are inherently explainable, including decision trees, decision rules and linear regression models. But these algorithms are best-suited for simpler problems, as they trade off power for explainability.

3. More powerful AI algorithms can be approximately explained using a growing body of sophisticated techniques, including Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

Understanding AI Explainability

AI explainability is about finding ways to make the causes and meanings of AI behaviours more transparent to people who interact with AI systems. The techniques for doing this can be usefully

divided into two groups: **intrinsic** explainability techniques and **post-hoc** explainability techniques.

Intrinsic Explainability

Intrinsic explainability arises when AI systems use algorithms that are inherently transparent and understandable without any need for additional techniques to supplement the algorithm. Intrinsic explainability is possible because not all AI algorithms are naturally opaque and difficult to understand. AI opacity is especially associated with modern deep learning algorithms and neural networks. These systems work by sending data through hundreds of hidden layers that perform thousands or millions of calculations to transform the input into the desired output. The details and meanings of these inner computations are not known even by human programmers of AI systems, making it very difficult to give a humanly-understandable explanation of the system's behaviour in any given situation.

By contrast, intrinsically explainable AI algorithms are more similar to traditional software programs in that they use procedures that can be broken down into a series of rule-like decisions.

Intrinsic explainability is best suited for when you have more straightforward data sets that simpler models can learn. Intrinsic techniques are great for transparency but are not as powerful or accurate as modern deep learning techniques.

Post-Hoc Explainability

Post-hoc explainability techniques are applied to the behaviour of opaque AI systems after they are trained and used to generate outputs. In other words, the explainability techniques are not baked into the AI algorithm itself, as with intrinsic explainability techniques, but are applied after the fact.

Post-hoc methods have the advantage of being widely applicable to many different AI algorithms. Using a post-hoc technique means not limiting yourself to simpler and less powerful intrinsically explainable algorithms. The drawback is that the insights generated by post-hoc methods are generally approximations that are not as accurate as the explanations available for intrinsically explainable algorithms.

The Importance of AI Explainability

A strong business case can be made on several grounds for emphasising explainability when deciding which AI systems to use and develop.

Building Trust

An explainable AI system is a trustworthy system. Users and other stakeholders are more confident in the decisions and outputs of AI when they can understand its behaviour.

Transparency also builds a framework for regulatory trust by bringing to light the factors that shape AI decisions.

Ensuring AI Accountability

Without an understanding of why an AI system makes the decisions it does, there cannot be a detailed accounting when errors occur. Opaque AI systems can also lead to human complacency. If a human user or developer cannot peer into the black box of AI decision-making, they will be less effective in scrutinising the outputs and decisions of that process.

Facilitating Regulatory Compliance

Governments and regulatory agencies around the world are racing to put in place legal frameworks that govern the development and use of AI systems. Already, the E.U. AI Act passed the European Parliament, and U.S. President Biden issued an executive order mandating new standards for AI safety and transparency. There are already statutes mandating some degree of explainability in the use of AI algorithms. For example, the General Data Protection Regulation in the E.U. specifies that when businesses possess personal data about consumers and use "automated decision-making," consumers have the right to "meaningful information about the logic involved" in that automated decision. It is an almost sure bet that as the regulatory environment around AI grows larger, there will be more requirements on businesses related to AI explainability and transparency.

Techniques for AI Explainability

The following are some specific techniques used in AI explainability. To understand these techniques, it is important to grasp the concept of a feature. A feature is a data point used to represent some characteristic of the thing that the AI model learns about. For example, an AI model trained on images of trees might have features related to the tree width, the density of leaves, and so on. Specific values for features are used by AI systems to classify input data.

Intrinsic Methods

Some AI algorithms are inherently explainable. These intrinsic methods include:

- **Decision Trees**: Decision trees work by creating cutoff values for features that are used to sort data inputs into groups known as "terminal nodes." A series of cutoff decisions is made before the data gets to these terminal nodes. For example, there might be a cutoff for trees taller than 20 metres, then another cutoff for trees with needle-shaped leaves, and so on. Once the data is sorted, an averaging technique is used to determine the final output. Decision trees make interpretability straightforward by being able to identify important cutoff values for features and being able to trace the path of an input through the tree.
- **Decision Rules**: Decision rules work by creating simple IF-THEN statements to make decisions. For example, an AI system might have a rule that IF a tree has a certain height

and has bark of a certain color, THEN that sample gets sorted into a particular group. Once training is completed, the model will have a number of these IF-THEN rules for classifying data. One can interpret the model easily by examining these rules, which give a complete picture of the model's inner logic.

- **Linear Regression**: Linear regression models use linear equations to transform the input features into output decisions. Since the relationship between the features and the output is linear, it is straightforward to determine how much each feature affects the final determination. For example, you would be able to say how the output would have changed had the tree height been different by five metres.

Local Model-Agnostic Methods

Unlike the previous methods, local model-agnostic methods do not require a specific AI algorithm to work. Instead, they can be applied to any AI model post-hoc. They are also "local" in the sense that they are used to explain individual decisions. When these methods work, in other words, they can tell you why an AI system behaved a certain way in a specific instance.

- **Local Interpretable Model-Agnostic Explanations (LIME)**: LIME is a technique that creates a surrogate model of an AI system using one of the intrinsic methods above. This surrogate approximates the behaviour of the opaque AI system and allows you to explain its individual decisions. Essentially, a simpler version of the opaque system is created and studied to understand its behaviour.
- **SHapley Additive exPlanations (SHAP)**: SHAP calculates a measure of the importance of each feature using tools from game theory. Each feature is thought of as a player in a game to optimise the accuracy of AI decisions. These techniques are used to create a linear model of the relationship between features and decisions, allowing for an approximation of the linear regression method for explainability described previously.

Global Model-Agnostic Methods

The following methods are model-agnostic methods that try to describe the behaviour of an AI system as a whole, rather than explaining particular decisions made by the system:

- **Feature Importance Analysis**: Feature importance measures the amount of impact that each feature has on the final decisions of a model. There are a variety of methods you can use to calculate feature importance, and each of the techniques described above allows for analysis of feature importance. This analysis lets you determine which characteristics are most important to the final behaviour of the system, increasing understanding of the model.
- **Visualisation Tools**: Many people find information easier to absorb when it is presented visually. You can use a variety of visual tools to illustrate the behaviour of AI systems, including graphs representing feature importance. Decision rules and decision trees are especially well-suited to visualisation.

Best Practices for AI Explainability

Implementing these practices will aid your AI explainability efforts.

Integrate Explainability From the Start

It's important to consider explainability right up front if you're developing an AI system. You should consider whether the task and data are simple enough that you could successfully use an intrinsically explainable model to get the job done. Only use an opaque model design if it is necessary to achieve satisfactory performance.

Tailor Explanations to Stakeholders

AI explainability is about giving helpful explanations, but what counts as helpful depends on the context and the audience. For non-technical audiences, you should aim to give simpler explanations like basic visual illustrations or a list of the most highly relevant features. Technical audiences like AI engineers themselves will want more details, such as the calculation of Shapley values.

Continuous Monitoring and Improvement

The field of AI explainability is still young and developing. You should continuously monitor not only your own organisation's use of explainability techniques, but the development of new techniques and methods in the broader explainability community to ensure that your methods are up to date.

Challenges in AI Explainability

Explainable AI techniques have not yet been perfected. The most fundamental challenge lies in the trade-off between the power and capability of AI systems versus their interpretability. As discussed, simpler algorithms are highly transparent, but not as powerful as modern deep learning. While surrogate models (such as LIME) and other techniques can somewhat bridge this gap, the trade-off still exists. Further interdisciplinary work will be needed to fully cross the divide between human- and machine-understandable models.

Final Thoughts

Transparency and explainability increases our trust in — and control over — complicated AI systems. In the developing regulatory environment around AI, explainability will be both an ethical responsibility and a legal requirement. While the field of AI explainability is still developing, there are already several powerful techniques to help you crack open the black box of AI technology. Zendata can help you chart a course in AI governance that makes your systems explainable.