THE NEED FOR STANDARDIZED EXPLAINABILITY

Othman Benchekroun[1], Adel Rahimi[1], Qini Zhang[1], Tetiana Kodliuk[1]

Abstract

Explainable AI (XAI) is paramount in industry-grade AI; however existing methods fail to address this necessity, in part due to a lack of standardisation of explainability methods. The purpose of this paper is to offer a perspective on the current state of the area of explainability, and to provide novel definitions for Explainability and Interpretability to begin standardising this area of research. To do so, we provide an overview of the literature on explainability, and of the existing methods that are already implemented. Finally, we offer a tentative taxonomy of the different explainability methods, opening the door to future research.

INTRODUCTION

Motivation

It is undeniable that we are living in the era of Artificial Intelligence (AI). However, there is a major issue with recent machine learning models: although they yield great results, all of them are ”black-box models”. This leads to Unintentional Misbehaviors, which poses an even greater danger as we may have errors that were not accounted for, but still affect millions of people.

Some biases inherited from the humans building these AI models were highly publicized, such as:

Gender Bias:
Job openings for top positions were only showed to men

Race Bias:
An automatic soap dispenser only recognized white users and not black users.

Confirmation Bias:
This is the biggest challenge facing modern societies due to the bubble effect of social media.

The need for XAI in the Industry

Organization have to use explainable AI

to comply with the privacy-related regulations all over the world putting the data subjects back in control of their personal information (e.g. GDPR in Europe, PDPA in Singapore, CCPA in California, …) All these laws force companies in some way to explain the decision-making process of their algorithms .

&

Organization want to use explainable AI

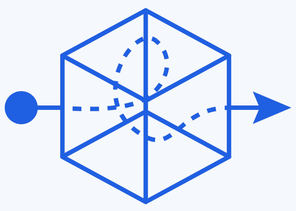
given all the advantages that explainability offers. Not only do organizations gain valuable insights on the behaviour of each of the models they use, they can also build more efficient, profitable and cost-wary AI models.

DEFINITIONS & RESULTS

Characteristics for Explainable AI

According to Merriam Webster, to explain is “to make something plain or understandable”. As such, we define 3 levels of explanation of AI models, which answer a particularly question posed by users:

1. Interpretability, or ”How does the AI model behave?”



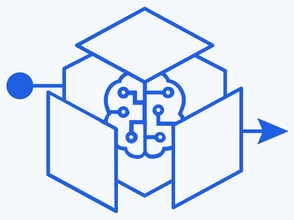
Definition 1. *The interpretability of AI models refers to the condition that allows a user to understand the relation- ship between the input and the output of the AI models, providing a clear grasp of all resulting data points.*

2. Auditability, or “Does the AI model behave as expected?”

We can define the most important property of an auditable AI models its capacity to answer specific questions auditors might have such as: *“Are there any bugs?”*, *“Are unintentional biases included in the model?”*, *“Does this model have se- curity risks?”* or even *“Is this AI model compliant with a specific data protection regulation?”*

3. Explainability, or “How is every decision taken by the AI model?”

Definition 2*. The explainability of AI models refers to the condition that allows a user to understand decisions made by the model and its subparts through processes before, during, and after the construction of the AI model.*



Some Explainability Methods

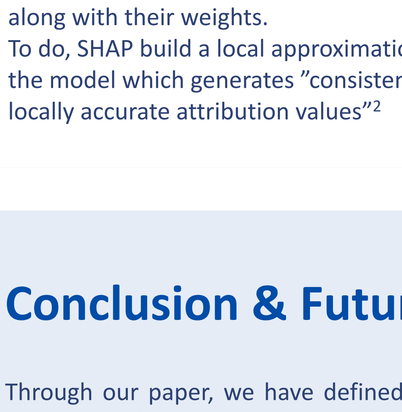


Figure 1: A plotted Decision Tree with its attributes and labels, trained on the Iris dataset

1. Inherently Explainable Models

Inherently explainable AI (IXAI) models are white- box models that are transparent by design and highlight the main features used for prediction.

- **Decision Trees** are very inexpensive and extremely fast to build.
- The main advantage of this architecture is that decision trees are highly interpretable white-boxes.
- However, they are not robust, and therefore extremely prone to overfitting.

2. Interpreting Black-Box Models

Unfortunately, not all problems are simple enough to be solved by IXAI models and require more complex models that aren’t so transparent. To provide explainability to these models, multiple methods were created since 2016

- **SHAP** or *Shapley Additive exPlanation*, is an approach based on Shapley values, used in game theory
- This method belong to the class of **additive feature attribution methods** where the result is a linear combination of features along with their weights.
- To do, SHAP build a local approximation of the model which generates ”consistent and locally accurate attribution values”[2]

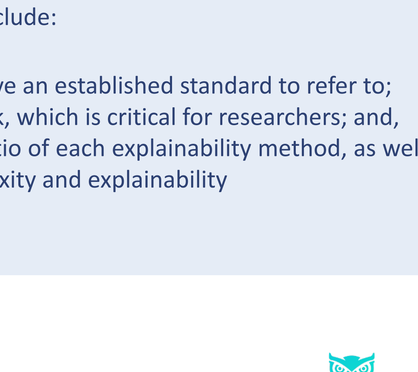
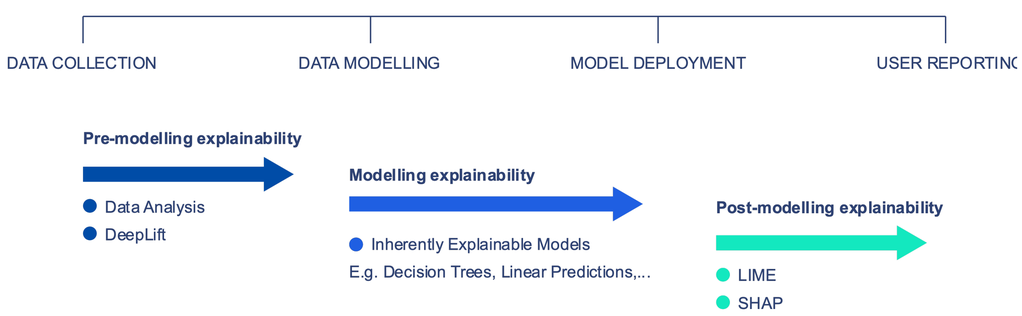


Figure 2: An example of Probability Visualization from SHAP

FINDINGS

Taxonomy of Explainability Methods



Conclusion & Future Works

Through our paper, we have defined the meaning of AI explainability and established its importance for the construction of trust-worthy and intentional Machine Learning models. This opens the door to a wider adoption of XAI and a systematization of this field; for if we do not drive explainability today, it will be too late tomorrow to correct the wrongs of AI and ensure a safe future where AI is omnipresent.

Other interesting leads to follow this research include:

- a more comprehensive taxonomy allowing to have an established standard to refer to;
- a standardisation of the explainability framework, which is critical for researchers; and,
- an analysis of the explainability to complexity ratio of each explainability method, as well as the possible links between performance, complexity and explainability

[1] Dathena Science Pte Ltd. Singapore, Singapore
[2] Lundberg, S. M. et al., Consistent individualized feature attribution for tree ensembles, 2018

dathena