EXPLAINABLE AI

Academic Year: 2023/2024

Author
Lorenzo Bonanni
lorenzo.bonanni@studenti.univr.it

October 10, 2023

Contents

Part I
Introduction & Motivation

AI (and particularly ML) can solve very complex problems in real life such as: Autonomous vehicles, Industrial automation and robotics Medical imaging and diagnosis ecc.. The problem with most of the ML models is that they are black boxs i.e we cannot understand the motivation behind the output.

The goal of Explainable AI (XAI) is to explain the behaviour of AI systems, in order to increase the level of understanding and *trust*.

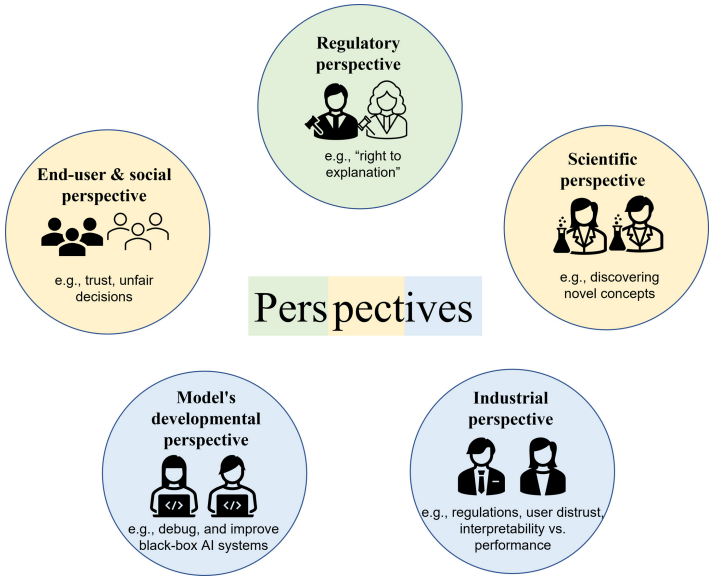

Figure 1: Summary of the main motivations regarding the use of XAI

We can summarize the motivation into 5 main points:

1. Regulatory Perspective
2. Scientific Perspective
3. Industrial Perspective
4. Model's developmental perspective
5. End-user and social perspective

1 Motivation

Regulatory Perspective

Black-box AI systems are being utilized in many areas of our daily lives, which could be resulting in unacceptable decisions, especially those that may lead to legal effects. The European Union's General Data Protection Regulation (GDPR)[1] is an example of why XAI is needed from a regulatory perspective. These regulations create what is called the

[1] `https://www.privacy-regulation.eu/en/r71.htm`

right to explanation, by which a user is entitled to request an explanation about the decision made by the algorithm that considerably influences them.

Scientific Perspective

When building black-box AI models, we aim to develop an approximate function to address the given problem. Therefore, after creating the black-box AI model, the created model represents the basis of knowledge, rather than the data. Based on that, XAI can be helpful to reveal the scientific knowledge extracted by the black-box AI models, which could lead to discovering novel concepts in various branches of science.

Industrial perspective

Regulations and user distrust in black-box AI systems represent challenges to the industry in applying complex and accurate black-box AI systems. Less accurate models that are more interpretable may be preferred in the industry because of regulation reasons. A major advantage of XAI is that it can help in mitigating the common trade-off between model interpretability and performance, thus meeting these common challenges. However, it can increase development and deployment costs.

Model's developmental perspective

Several reasons could contribute to inappropriate results for black-box AI systems, such as limited training data, biased training data, outliers, adversarial data, and model overfitting. Therefore, what black-box AI systems have learned and why they make decisions need to be understood, primarily when they affect humans lives. For that, the aim will be to use XAI to understand, debug, and improve the black-box AI system to enhance its robustness, increase safety and user trust, and minimize or prevent faulty behavior, bias, unfairness, and discrimination.

End-user and social perspective

In the literature of deep learning, it has been shown that altering an image such that humans cannot observe the change can lead the model in producing a wrong class label (Adversarial Attacks). On the contrary, completely unrecognizable images to humans can be recognizable with high confidence using DL models. Such findings could raise doubts about trusting such black-box AI models. The possibility to produce unfair decisions is another concern about black-box AI systems. This could happen in case black-box AI systems are developed using data that may exhibit human biases and prejudices. Therefore, producing explanations and enhancing the interpretability of the black-box AI systems will help in increasing trust because it will be possible to understand the rationale behind the model's decisions, and we can know if the system serves what it is designed for instead of what it was trained for. Furthermore, the demand for the fairness of black-box AI systems decisions, which cannot be ensured by error measures, often leads to the need for interpretable models.

2 XAI - One topic, two keywords

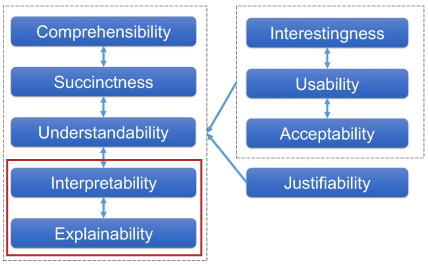

Figure 2: Outline of the relationships between the common XAI terminologies

In the Explainable AI domains there many terms that seems to be the same meaning. Specifically we are interested into two terms: Explainablility and Interpretability.

Explainablility provide insights about an AI model while Interpretability tells if the insights make sense to the audience.

3 Types of XAI

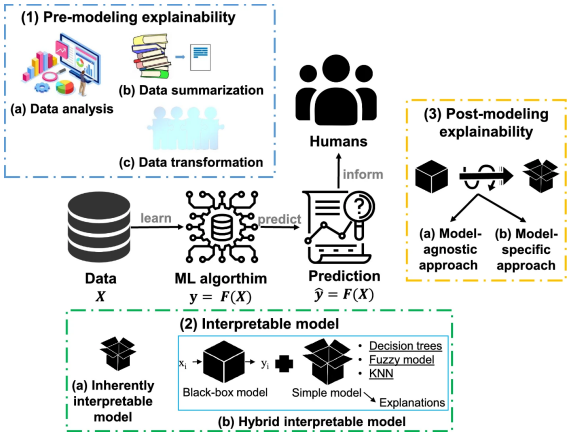

Figure 3: Outline of the relationships between the common XAI terminologies

There are three main types of Explainablility:

1. **Pre-modeling explainability**: summarize input data to identify most relevant features or aspects, based on statistical analysis. Some examples are: K-Means and PCA

2. **Post-modeling explainability**: explain the results of a black-box model. Some tecniques include:

   - *Feature relevance*: which input data mostly influences the output?
   - *Simplification*: aims to make a simplified version of the original model that has an optimized function, significantly reduces the complexity, has a simpler implementation process, and performance is comparable to the original version.
   - *Visualization*: interprets a models behavior by visual representations. Visualiza- tion techniques are considered the best way to explain the complicated inner interac- tions of the variables of the model, and they can be combined with other methods in order to increase their interpretability ability.
   - *Textual justifications*: explains a model by generating explanations in the form of text
   - *Contrastive explanations*: clarifies why an event occurred in contrast to another

3. **Interpretable models**: the model is not black-box on its own. some examples include: Linear or logistic regression and Decision trees

There are three main properties for interpretability:

- Algorithmic transparency: The model can be expressed as a set of known mathematical or logical relations
- Decomposability: The model can be decomposed in submodules, with clear indication of connections between them
- Simulatability: The model can be easily simulated by a human, given only any input