Explainable Artificial Intelligence

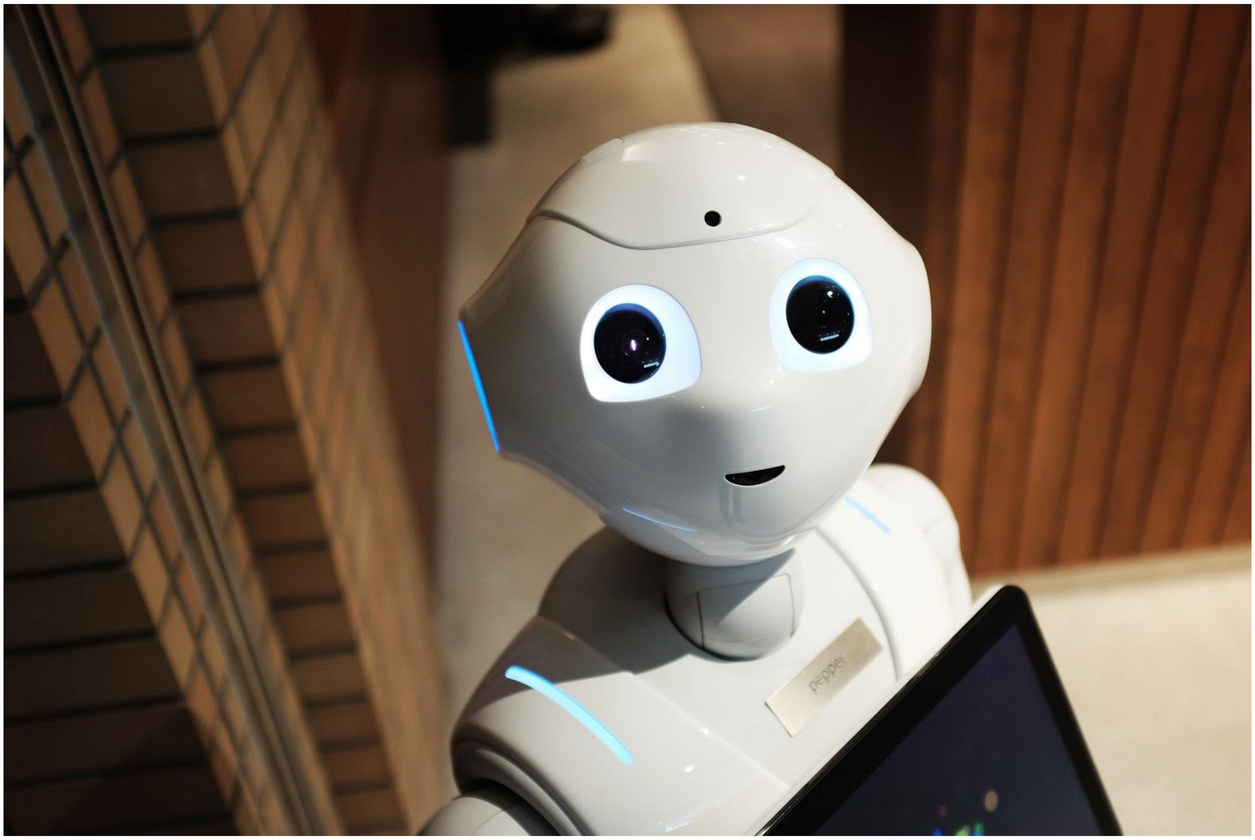

Photo by Alex Knight on Unsplash

Introduction

We have all seen how powerful and accurate an AI system could be, but it is so, at the cost of them being black box models as they lack the virtue of explainability. And hence, Explainability in AI has become a hot topic for researchers these days. Development in explainable AI is certainly very important for us humans to understand and become more trustful of the decisions a machine took for us.

Explainable AI (a.k.a. XAI) is a field that is rapidly emerging nowadays, it aims to explain how “black box” AI and machine learning models are actually able to make complicated decisions for us. It is inspecting and trying to understand all the steps and models that involve decision making. And thus, most of us (owners, operators and users) are expecting the XAI to answer hot questions like - Why a specific AI or machine learning model makes a certain prediction? When it would fail or succeed and with what probability it would be so? Or When can we really have enough confidence in the decisions made by a machine for us? And so on.

Explainability Vs Interpretability in AI

Before we go on any further first we need to understand the difference between the explainability and interpretability in AI.

In the machine learning context, explainability and interpretability are generally used interchangeably. But there exist subtle differences between them that might be worth knowing about.

Interpretability can be defined as the extent to which one can predict a model's result without trying to understand the reasons behind the predictions. So it can be said that it is easier to know the reason behind certain predictions or decisions if a model has higher interpretability. On the other hand, **explainability** is the extent to which the internal working of a machine learning or deep learning system could be explained to humans.

To understand the difference better, think of it like this - interpretability is about being able to know the mechanics of a model without being explicitly told. And explainability is the ability to clearly explain what is happening in a model and how does it work.

How to achieve explainability in AI models?

Explainability in AI, in a way, can be gained through using machine learning algorithms that are inherently explainable such as, Decision trees, Bayesian classifiers, and others that have both traceability and transparency, up to a certain level, in their decision making that could readily provide the visibility needed for AI systems without sacrificing too much of accuracy or performance. But these days we have more complicated and potentially more powerful models such as Neural networks, ensemble methods (e.g. random forests), and alike that, in general, sacrifice transparency and explainability for better accuracy, power and performance. Therefore, for explaining such complicated models we go for other more complex algorithmic approaches, that we will see below.

In order to develop explainable machine learning and deep learning systems, we can use any of the two main sets of techniques, i.e., **ante-hoc** or **post-hoc**. Ante-hoc are the set of techniques which introduce explainability into a model from the very beginning. Post-hoc techniques are just of the opposite of ante-hoc, that is they allow models to be created and trained normally and only at the testing time the explainability is introduced in the models.

Some Ante-Hoc Methods :

1) **Reversed Time Attention Model (RETAIN):** To help doctors understand the AI software's predictions, the researcher's at Georgia-Tech developed the RETAIN model. Data collected from many patients' hospital visits were sent to 2 RNN's having attention mechanism, which helped in explaining which part the neural network was focused on and which features actually influenced its choice.

2) **Bayesian Deep Learning (BDL):** We could measure the uncertainty of a neural network's prediction using BDL. By either learning a direct mapping to probabilistic outputs or by placing distributions over model weights the BDL forms uncertainty estimates. Using weight distributions of various predictions and classes, we can tell what feature led to what decisions and its relative importance.

Some Post-Hoc Methods:

1) **Local Interpretable Model-Agnostic Explanations (LIME):** Unlike RETAIN, LIME have a wide range of applications as it's not customized to a single domain. We can't call it a purely transparent model as it provides the explanation after a decision has

been made. For instance, in an Image classification problem (using CNN), we get the probability distribution over all classes. Then, by making small changes to the input we could see how it affect the distribution and collect the results. Then on collected data linear interpretable model is fit and the key features are extracted with their weights telling how prominent they are. It first blacks out different areas of the original image and feeds the resulting images into the model to see which of the new images through off the algorithm farthest and derive reasoning behind the model's predictions. For example, in an image of a tree frog, LIME showed that much of the classification decision was only based on the frog's face as erasing parts of the frog's face made it much harder for the model to identify it correctly.

2) **BETA:** BETA is considered to be closely related to interpretable decision sets. To explain the part of the model behaviour unambiguously, BETA learns a compact 2-level decision set. BETA uses an objective function which helps the learning process in having high fidelity (agreement between explanation and model), high interpretability and low unambiguity. All these aspects are combined into one objective function which is optimized.