Explanations and cognitive bias: an underspoken issue of the XAI approach

Alvise de' Faveri Tron
Politecnico di Milano
alvise.defaveri@mail.polimi.it

***Abstract*—**
TODO

I. Introduction

AI has made giant steps since its birth in the late '50s, especially in the last decade. Many tasks that in the past were exclusively carried out by humans, for example in the legal, law enforcement and medical fields, are now being automated with AI. This increase in AI's capabilities is a game-changer for today's technology and society, so much that expressions like "AI singularity" and "Fourth industrial revolution" have been used to describe this phenomenon.

However, today's AI is far from being perfect, and our knowledge of this technology is still partial and prone to misinterpretations.

AI tends to be unpredictable, and the nature of many AI algorithms is such that errors in their internal models can be difficult to identify and correct, remaining latent for potentially long periods of time. Machine Learning and Neural Networks in particular, which are central to many AI systems today, are difficult for humans to debug, and there's no such thing as "fixing one line of code" in these systems.

This is known as the *black box problem* in AI, and it poses huge ethical and practical concerns about whether we can trust this technology or not, especially given the nature of the tasks we expect them to be able to undertake in the future.

Explainable Artificial Intelligence (XAI) is a recent field of AI that aims at addressing the problem of understanding AI and making it more reliable and trustworthy. Many interesting results have been achieved in this field, and many more are expected to come in the next years. However, there are still some fundamental aspects that this approach seems to struggle with, for example the problem of giving a formal definition of explainability, and others that cannot be solved within the XAI framework alone, for example the problem of distinguishing correlation from causation.

In this paper we want to focus on the problems that arise when using an AI explanation interface as a proxy for measuring the quality of an AI algorithm, and reason about the biases that the explanation itself introduces in this measurement.

We propose a number of metrics over which an explanation can be characterized. In particular we focus on *fidelity*, i.e. the adherence of the explanation to the original system, versus *complexity*, which is the metric that many XAI studies use to evaluate their solutions. Our goal is to show that a tradeoff between these two metrics exists, and that this introduces the possibility for cognitive biases to play a role in evaluating an AI output.

In this context, we will use the words "explainability" and "interpretability" interchangeably, as suggested in .

The reminder of this paper is organized as follows.

Section II provides some examples that we consider relevant to understand the XAI problem in detail.

Section III gives a more specific definition of XAI and a general overview of the solutions that are being developed in this framework.

Section IV tries to break down the concept of explainability in several dimensions, which can be used to identify and classify AI explanations.

Finally, in Section V we highlight the problem of measuring fidelity and quality in an explanation, and show which cognitive biases could be introduced when examining an AI trough an explanation interface.

II. Background

A. AI today

The term "Artificial Intelligence" has historically been used with many different meanings, and this is especially true today that this subject is receiving a lot of media attention. While giving a definition of AI is out of the scope of this paper, in this Section we just want to highlight some aspects of modern AI that motivated the research for a more explainable AI.

First of all, AI algorithms are generally *meta*-algorithms, which means that they provide a recipe to create algorithms that can explore the solution space of a problem in an "intelligent" way, whatever meaning might be associated to this term. This means that the same AI algorithm can be used to solve very different problems, such as image recognition and stock market trading strategies.

Secondly, many modern AI techniques, such as Neural Networks, Genetic Algorithms, Swarm Intelligence etc., have some kind of non-determinism embedded in them: this is due to the fact that in real-world problems typically there is no optimal solution to find, but rather many suboptimal solutions that can be computed in a reasonable time, and adding randomness typically speeds up the process of finding candidate solutions by orders of magnitude.

Another aspect of many modern AIs is *learning*. Machine Learning techniques generally consist in feeding an algorithm

with a large number of instances of a given problem and letting it figure out on its own the best way to model it. The strength of this technique is that no previous knowledge of the model of the problem is needed, nor it can be enforced generally, and this is exactly where it gets an edge over more traditional computer science approaches.

A representation of this idea is provided in Figure 1

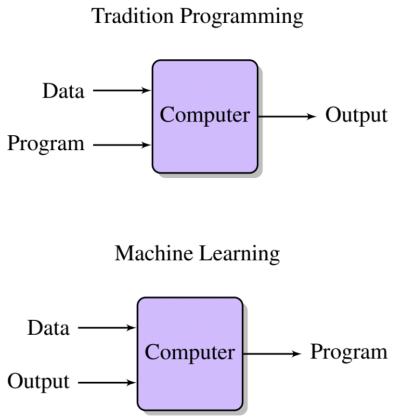

Fig. 1. Difference between traditional computing an machine learning: while in the former a program is provided to the machine, in the latter it is the machine itself that is able to synthesize it.

These aspects together mean that even the developer of an AI algorithm is not fully in control of all the decisions that the AI system makes, which is one of the main reasons why AI is considered to be a *black box*.

B. Explaining AI: a toy example

As an example of the problem of explaining an AI system, we can take a look to what it is like to work on a Neural Network. A concept representation of a Neural Network's structure is depicted in Figure 2.

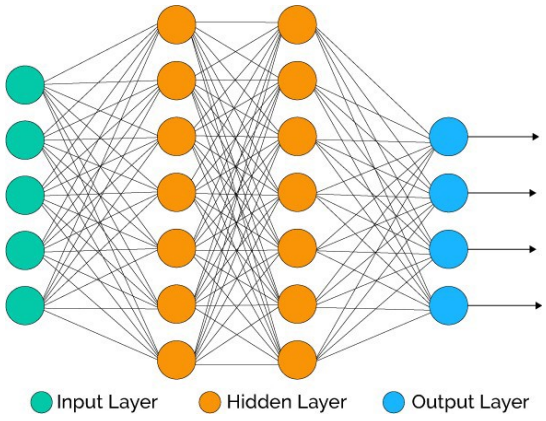

Fig. 2. Simple representation of an Artificial Neural Network

The graphical representation is useful for understanding the general architecture, but it doesn't really tell us anything about how the Neural Network actually works, i.e. what is the relationship between a certain input and its output.

We could give a more precise representation of this dependency in Figure 3, which explicitly defines the mathematical relationship between the input and the output. Without going into the details of the mathematical formula, we can see how we would still have a hard time understanding what a Neural Network does if we were to adopt this representation.

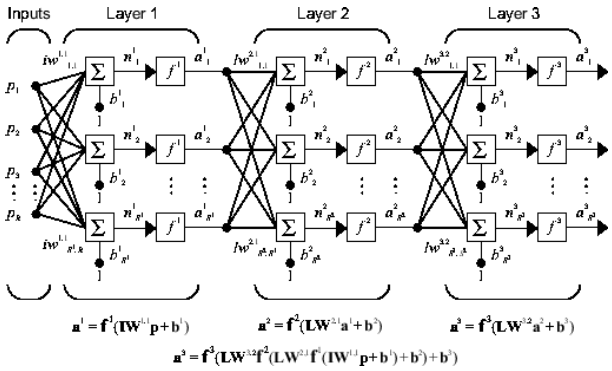

Fig. 3. Mathematical equivalent representation

On the other hand, Figure 4 represents a *Decision Tree*, another family of AI algorithms.

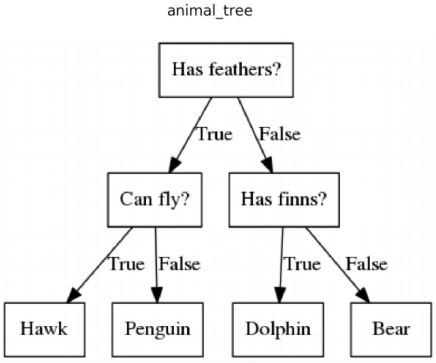

Fig. 4. A simple Decision Tree

There is a clear difference with the previous example: by looking at the structure of the algorithm we can immediately be sure of what decisions it is going to take, and get a grasp on the chain of reasons that caused a specific output. The main issue of these algorithms is that, for complex problems, they tend to be outperformed by Neural Networks, and the efficient variations of these techniques such as Random Forest greatly increase the complexity of the decision trees, which can grow very big and hence be very hard to understand.

This simple example shows how the solution space to the explainability problem has multiple dimensions, constraints and trade-offs that have to be taken into account.

C. Bias problems in AI

Data bias in AI, demonstrate how vulnerable is AI to biases

III. The XAI approach

A. The goal

Explainable AI is a concept that was recently formalized in a call for papers made by DARPA, the same agency where the term "Artificial Intelligence" was born in the first place. It is meant to be describe a new set of Artificial Intelligence systems which are designed to be easier to understand by humans. In particular, the goal of XAI is to make Artificial Intelligence more:

- **Easy to debug** for the developers
- **Predictable**, so that companies and governments adopting this technology can be aware of the possible weaknesses of a their models, and can be held responsible when using bad algorithms
- **Trustful** for operators and end-users of this technology

Figure 5 describes the end result that is expected from XAI.

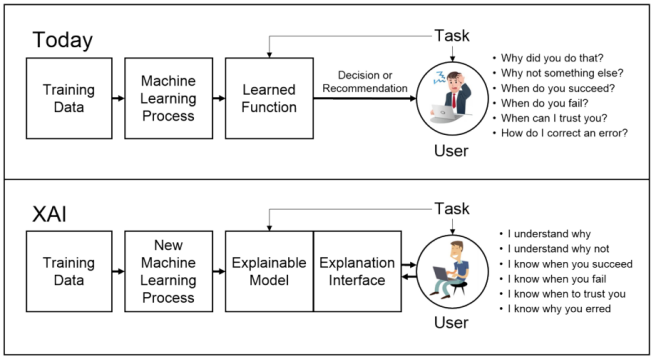

Fig. 5. The goal of XAI as expressed in the official DARPA presentation of the XAI project.

The creation of XAI requires the join effort in a variety of research fields, from Computer Science to Cognitive Psychology, and there is still a lot of work to do. Nevertheless already many papers have been submitted on the subject, indicating a growing interest of the research community towards this subject.

B. Current Solutions

Given the highly experimental nature of this topic, many different solutions have been proposed by various papers in the framework of XAI, which vary greatly in intended use, goal and adopted approach. Giannotti et al. contains a classification of the existing XAI solutions and their respective strengths and weaknesses. With no aim of being exhaustive, here we give an abstract classification of the main XAI solutions, based their general approach to the problem.

XAI approaches might be classified as:

1) **Visualization**: improving the understanding of a model using a better way to visualize its internals. One popular application of this approach is computer vision, where features of the learned model might be mapped onto the input image.
2) **Simplification**: similarly to the idea explained in Section II-B, this approach consists in trying to adopt simpler models or simplify already existing models to just a set of important features.
3) **Approximation**: once a model has been produced by an AI, one explanation technique is to try and understand the dependencies between inputs and outputs by trying to find which output change is triggered by a given input change. Most of the times this means, in practice, creating a behavioral model of the algorithm which is parallel to the algorithm itself, and has no immediate correlation with the algorithm's internal structure.
4) **Explanation by Example**: a nice information to have when trying to understand an AI model, especially in the case of classifiers, is an example, or a *prototype*, of how the AI thinks that a typical member of a given class should appear. This can be realized in many ways, for example by attaching to a classification output a set of minimal changes to the input that would cause the output to be modified, or specify a partially filled object for each class.

Immagini esemplificative

Si può integrare con una classificazione più classica delle tecniche (internal vs external ecc)

IV. Defining Explainability

As we anticipated in Section I, one fundamental problem in the field of XAI is that there is no single conventional notion of explainability.

If, on one hand, many solutions have already been proposed to tackle the problem, with various claims regarding their interpretability, on the other hand the lack of a formal definition seriously challenges the findings of these researches, casting a shade of doubt on the proposed solutions.

Mythosgoes as far as considering the term itself ill-defined, therefore stating that claims about interpretability generally have a quasi-scientific nature. Giannottion the other hand, in a review of the current state of the art, considers the lack of a mathematical description as an obstacle for the future development of this field. The DARPA paper itself defines the formalization of an evaluation metric for explanations as one of the goals of the XAI project, to be developed in parallel with technical solutions.

Without discouraging the research on this matter, we want to highlight that this is easier said than done.

A. What is the scope?

Before evaluating an explanation interface or an XAI system in general, we should ask ourselves at least the following questions:

Explainable to whom? The concept of *user* of an AI system is not always well defined, nor is the concept of user of an explanation. This might include:

- The **developer** of the AI system, as he is only partially in control of what the algorithm does (refer to Section II-A)
- The **operator** of an AI system: many AI algorithms nowadays are being used as an input for a human to make decisions on a certain subject
- The **end user** which is affected by the decision of an AI

Explainable for which purpose? Different users have different needs, that can partially overlap, when it comes to AI explanation. More in general, whether a certain representation can be considered explanatory depends to some degree on what it is being used for. In the case of XAI, some common purposes are:

- **Debugging**: finding errors or underperforming portions of the system
- **Human-in-the-loop**: creating systems where human and AI decisions can co-exist and influence each other
- **Validation**: understanding if a certain model is good enough to be deployed for a certain tasks, where it fails and what happens when it fails
- **Appeal AI decisions**: [1] giving the right to users and citizens that are affected by AI decisions to know, understand and possibly appeal decisions that are automated with AI systems

It appears quite evident that different XAI solutions with different scopes and intended users cannot be compared in the same way.

B. Possible Metrics

Bearing in mind the different goals that an XAI system can have, we can identify a series of characteristics that are different among different solutions:

- **Complexity**: how many elements are there in the explanation?
- **Clearness**: how cognitively hard is the explanation? How difficult is it to understand the correspondence between the elements of the explanation and the information we are trying to gain?
- **Informativeness**: how much information, weighted on how meaningful is is, can be extracted by the explanation? E.g. does the explanation significantly modify the level of uncertainty about the AI behavior?
- **Fidelity**: how closely does the explanation represent the internal functioning of the system? Are all the facts inferred from the explanation also applicable to the original system?

Clearly, the choice of the evaluation metric

C. Can we measure all of them?

fidelity vs complexity

[1]This last goal is not explicitly listed in the original scope of XAI goals, but has gained traction recently with the publication of *right for an explanation* law in EU.

cit

D. Possible biases

cognitive bias

V. Some fundamental issues

- causality vs correlation
- previous knowledge
- ethical implications

VI. Conclusions

In conclusion, the main problem of XAI is that there is no single definition of what an explanation is, it depends on the purpose and on the user of the AI system.

For this reason, these should be considered different problems, at least the debugging problem vs the right of explanation problem: they are not correlated and saying that one solves the other poses some threats on the quality of the result itself.

"I always thought something was fundamentally wrong with the universe"

[1]

References

[1] D. Adams. *The Hitchhiker's Guide to the Galaxy*. San Val, 1995.