Can AI be explained? Some inherent limitations of XAI

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

***Abstract*—**
TODO

I. Introduction

Artificial Intelligence has come a long way since its birth in the late '50s. cit? In the last decade, in particular, we have seen amazing improvements in this field.

It has become increasingly pervasive, with successful applications in the medical, legal, law enforcement, financial and automotive fields. This success is undoubtedly shaping today's and tomorrow's society, since many tasks that were exclusively carried out by humans can now be handled in an autonomous way.

However, AI as a technology has just started its journey: there are currently many limitations and shortcomings to this technology, that emerge when observing its real-world implementations. One issue in particular has been found to be recurring and problematic, since it deeply undermines the trust and reliability of this technology: the fact that we don't understand enough about it.

Machine Learning and Neural Networks in particular, which are the most common and performing AI algorithms today, are difficult for humans to debug. Introspection tooling around this kind of technology is currently insufficient, and yet essential. We need to devise new methods to validate AI algorithms, find weaknesses and possibly correct them with a human-in-the-loop approach.

XAI is a relatively new field of AI that aims at addressing this problem. Many interesting results have been achieved in this field, and many more are expected in the next years, but there are still some fundamental issues, which are inherent in the AI explainability problem, that have not been addressed in a general and consistent way.

The goal of this paper is to identify and discuss some of such inherent issues, analyze some of their root causes, and argue that the current XAI approach generally fails to address them.

A. Structure of this paper

Before we discuss the problems of XAI, we need to give some context about what XAI is trying to achieve and for which purpose.

In Section II we will start by looking at some examples that should clarify the problem space.

In Section IV we give a more specific definition of XAI and a general overview of the solutions that are being developed in this framework.

Section III tries to analyze some of the root causes of the problems that are being tackled by XAI, giving a classification of various types of explanations.

Finally Section V contains a reflection on what is not being considered by XAI, in particular: the causality vs correlation problem, the problem of measuring the quality of explanations, the biases introduced by an explanation, and the problem of measuring the fidelity of an explanation.

II. Problem Overview

A. Current AI shortcomings

One very popular and effective type of Machine Learning techniques in use today is Supervised Learning. This approach generally consists in *training* an algorithm by giving it as input a large number of instances of a given problem that we want to solve, e.g. a classification problem or some kind of prediction, and letting it figure out on its own how to rearrange its internals in a way that is suitable to output the expected results. The strength of these systems, and of Machine Learning in general, is that no previous knowledge of the model of the problem is needed, or should be enforced for that matter, and this is exactly where this technology gets an edge over more traditional computer science approaches.

A general observation of how these systems behave in various fields have shown us one interesting fact: when these systems break, they tend to break hard. A misprediction made by an AI, especially if it has to accomplish a highly complex or impacting task, can cast a shade on the correctness of the whole model itself, on the data it has been trained on or on its design. There's rarely such thing as "fixing one line of code" on deep neural networks that have been trained on millions of data points: once its trained, you either add more data or start again from scratch, which can be a very high price to pay in terms of time and computational power.

Moreover, these kind of errors are generally difficult to predict in advance: an AI algorithm can perform very well on a high number of inputs, but have a weak point that is only discovered way after the AI has been deployed. There is no general method to know in advance where an AI might fail, we just know that until now it has been pretty accurate, which is one of the problems that XAI tries to address. As for today, we have very little introspection tools when examining

an AI system, especially when we are talking about neural networks: the same people that design and train the algorithm have generally little knowledge about what model the network is going to produce at the end, and when it does the only way of verifying its correctness is black box testing, for which the input space is generally huge.

All these considerations have encouraged the AI industry and the governments to tackle that which seems the hugest obstacle for AI being adopted everywhere: the problem of understanding an AI model and ”opening” the black box.

B. ”Explaining” a Neural Network

Of all the approaches that have been tried during its 50-years history, one has recently emerged as capable of solving many huge, complex and unrelated problems all together: Machine Learning. This particular form of AI is aimed at building a model of a problem from observing many examples of it, and letting the algorithm figure out the best way of connecting inputs and outputs, shaping itself based on what it ”learns” from the data. More specifically, one of the most studied and heavily developed approaches right now is Neural Networks, which encodes the information that has been learned as weights in the connections between basic units, called neurons.

A conceptual example of what a neural network is is shown in figure ...

rappresentazione grafica

However, this is just a *conceptual* representation. A more accurate representation of what is going on is a series of computations in the form of:

rappresentazione matematica

Which in itself is also an abstraction, since we don’t see here how the hardware is effectively calculating and solving these equations.

On the surface, this seems like a technical detail, but this is already the core of the AI Explainability problem: looking at these figures we are trying to get an idea of the *architecture* of the algorithm, but we would have a really hard time if we wanted to correlate what we see with the decisions that are being made by the algorithm.

III. Dimensions of the problem

A. Explainable to whom?

The users of an AI systems are:

- end user
- developer
- operator
- judge

B. Explainable for which purpose?

The main purpose for XAI are:

- debugging of the internals
- human-in-the-loop
- validation and certification
- appeal decisions

C. A multidimensional solution space

- complexity
- clearness
- informativeness
- fidelity

IV. The XAI approach

A. XAI: a new frontier of AI research

Explainable AI is a concept that was recently formalized in a call for research made by the DARPA, the same agency where cit the word ”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 goals of XAI is making artificial intelligence more:

- Debuggable
- Predictable
- Trustful

immagine presa dal paper del DARPA

This effort requires a variety of expertise, from Computer Science to Cognitive Psychology, and there is still a lot of work to do.

Fig. 1. The Universe

B. Solution Approaches

1) visualization
2) simplification
3) reverse engineering/fuzzing (exciting inputs with black box testing)
4) prototypes
5) differential for classifiers (what do I have to change to change the outcome)

V. What’s missing?

Ma come valutare le soluzioni? Basta la complessità?

The problem is that there is a lack of a formal definition of how an explanation should be measured. This is not a trivial point, since it is very difficult to quantitatively measure the goodness of any explanation. Many papers refer to complexity but this is not enough.

A. *Why is there an explainability problem in the first place?*

- causality vs correlation
- previous categories
- having a goal -¿ the decision means that I have to do something in practice, and that thing has an ethical an practical impact

B. *Can they be measured separately?*

- fidelity vs complexity
- fidelity vs clearness
- quality vs performance

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"

References