Measuring fidelity in AI explanations: an open problem of the XAI field

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

***Abstract*—Artificial Intelligence is known to suffer from interpretability problems: it is often hard for humans to explain why an AI took a particular decision or to understand what is the general logic behind its behavior. This is a serious ethical and technical barrier for the progress of AI, since this opaqueness has led, in some cases, to the realization of AI algorithms that behave in weird, unfair or dangerous ways. XAI (short for eXplainable AI) is a new research field that aims at solving this problem by providing some kind of explanation interface between humans and AI algorithms. However, at the moment, solutions in this field are being proposed without a widely accepted, common definition of *interpretability*. A common metric used for evaluating the interpretability of XAI solutions is complexity, for example the number of leafs in a decision tree: this is clearly not enough, since many other aspects, such as the usefulness and the preciseness of the explanations, are not taken into account. In particular, a commonly overlooked feature of explanations is *fidelity*, i.e. how closely does the explanation resemble the original system. This feature is generally opposed to simplicity, and the fact that many studies fail to explicitly measure this tradeoff leads to the following questions: how can we trust an explanation interface if we can't measure how accurately it represents the original AI system? Can inaccurate explanations introduce biases themselves by hiding or distorting aspects of the original model? If this is the case, isn't XAI introducing new trust issues instead of solving the existing ones that are inherent to AI?**

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 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: many decision systems based on AI still fail at tasks that are considered easy for humans, such as identifying objects in images or extracting salient information from text, and our understanding of this technology is still partial and prone to misunderstandings.

One well known issue of AI-based decision systems today is that they are difficult to debug, and many aspects of their behavior are not in direct control of the developer. Even when a decision system behaves well in a set of test cases, it is difficult to understand if the internal model correctly reflects the intended one. Machine Learning applications in particular tend to suffer from biases that are difficult to spot during the test phase, and they tend to display an overall opaque behavior which is not easily understandable for humans: 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. One of them is the problem of giving a formal definition of interpretability, which is a quite discussed topic in the field.

This paper focuses on some of the consequences of the lack of such definition, for example the fact that many papers in this field concentrate on the explanation's *complexity* without giving a precise evaluation of *fidelity*, i.e. the adherence of the explanation to the original system

This can cause situations where an explanation interface is too simple and might cover those aspects that would be otherwise identified as errors in its reasoning. Taken to the extreme, a system where the fidelity of an explanation interface is never taken into account can lead to a preference for those systems that can cover their defects better than others with respect to the specific explanation interface in use.

The final goal of this paper is to offer a reflection on how complex the evaluation process for XAI systems can be, and argue that XAI in general might be creating new trust problems that need to be explored by the research community.

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. Why is AI a black box?

The term "Artificial Intelligence" has historically been used cit 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.

cit 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 might be used to solve very different problems, such as image recognition, natural language processing 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.

Backup claims o riformulare

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

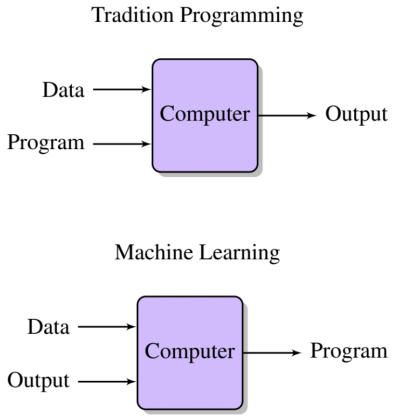

Fig. 1. Difference between traditional computing and machine learning: while in the former a program is provided to the machine, in the latter the machine itself is able to synthesize it.

Randomness and learning cause the developer of an AI algorithm to not be fully in control of all the decisions that the AI system makes, which motivates the idea that AI behaves as 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 conceptual representation of a Neural Network's structure is depicted in Figure 2.

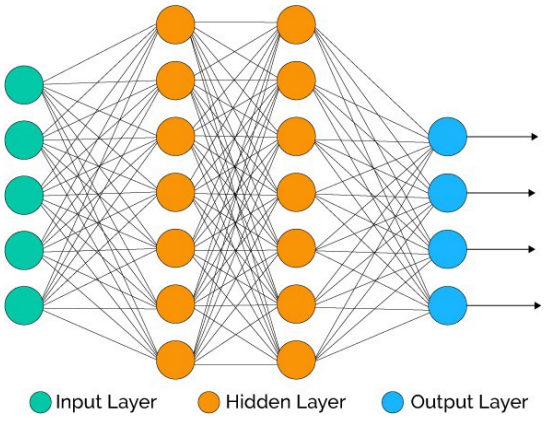

Fig. 2. Conceptual representation of the layers of an Artificial Neural Network

This graphical representation is useful for understanding the general architecture of the algorithm, 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. Although being more precise, we can see how we would still have a hard time understanding what a Neural Network does if we were to adopt this representation, especially in Deep Neural Networks with lots of intermediate layers between the input and the output.

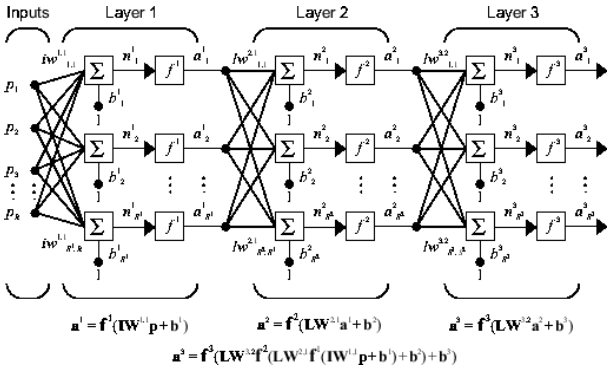

Fig. 3. Mathematical equivalent representation

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

The problem of hidden bias in AI is a well-known issue of current AI.

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

III. The XAI approach

A. Goals

Explainable AI is a new research line that was proposed by DARPA cit, the same agency where the term "Artificial Intelligence" was born in the first place cit, in anno. It is meant to be give birth to a new generation 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
- *Trustful* for operators and end-users of this technology

Figure 4 describes the end result that is expected from XAI.

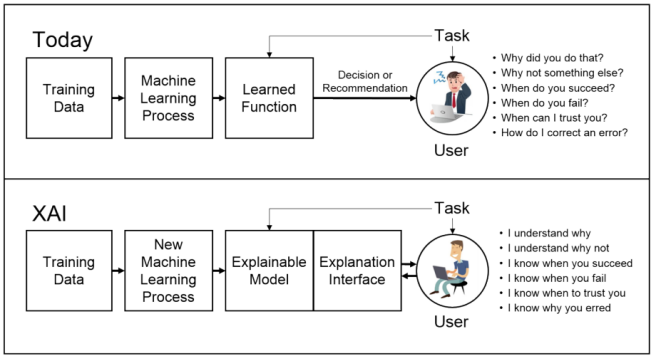

Fig. 4. The goal of XAI as expressed in the official DARPA presentation of the XAI project.
cit

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. cit contains a classification of the existing XAI solutions and their respective strengths and weaknesses, which are classified as shown in Figure 5.

The main techniques identified are:

- *Model explanation*: Global explanation of the whole model's behavior.
- *Outcome explanation*: Local explanation of a single outcome.
- *Model inspection*: Explanation trough introspection of the model's internals.

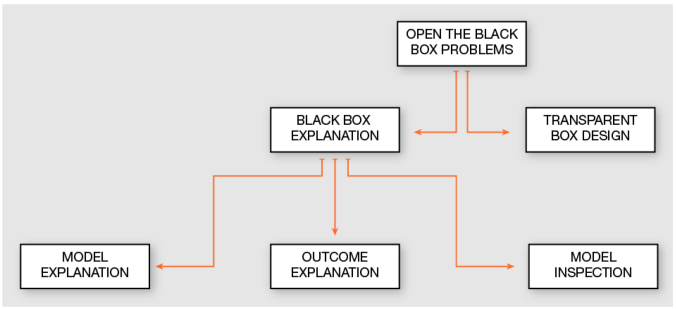

Fig. 5. Taxonomy of XAI solutions provided by
cit

- *Transparent Box design*: Making AI easier to understand by design, e.g. by using algorithms that are intrinsically clearer for humans.

With no aim of being exhaustive, we want to provide an abstract classification of the main XAI solutions, based their general approach to the problem.

XAI approaches might be classified as:

1) **Visualization**: these approaches focus on finding an effective way representation to visually represent key aspects of the AI behavior. One example is to use *saliency masks*, such as the one in Figure 6, to highlight which are the most significant portions of the input for the AI model, or in general to visually express which are the most important features that the model is recognizing. This is generally good for applications which are based on images, but might not be enough in other situations.

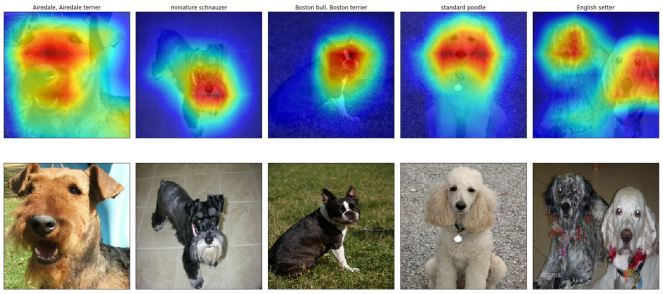

Fig. 6. An example of a heatmap representing the importance of pixels in a given picture, from blue (low importance) to red (high importance).
cit

2) **Approximation**: this approach consists in trying to adopt models that are simpler by design, or simplifying already existing models to just a set of important features. This is the case of *single tree approximation* cit, where the internal structure of an AI algorithm is approximated to a singe classification tree, such as in Figure 7.
3) **Behavioral Model**: this explanation technique focuses on expressing the dependencies between inputs and outputs ignoring the internal structure. This means, in

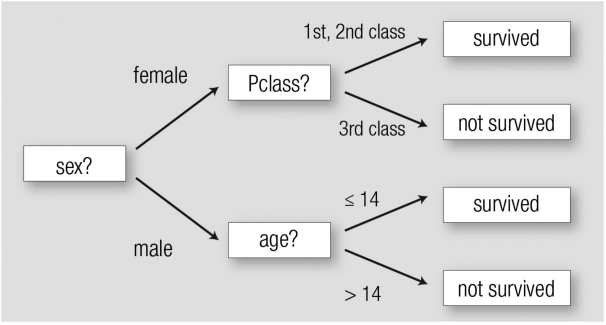

Fig. 7. A simple Decision Tree

practice, creating a behavioral model of the algorithm which is parallel to the algorithm itself, and can be expressed in an easier form, such as a set of logical rules.

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.

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.

cit 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 cit 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 cit 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.

[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

V. The troubles of explanations

A. Measuring fidelity

B. Fidelity vs Complexity tradeoff

C. How can an explanation introduce bias?

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.