

Are we teaching machines how to lie? The potential impact of cognitive bias in AI explanation

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-based system took a particular decision or to understand what is the general logic behind its behavior. This issue poses a serious ethical and technical barrier to the progress of AI, since the opacity of these algorithms has led, in some cases, to systems that behave in a weird, unfair or dangerous way. XAI (short for eXplainable AI) is a research field that aims at solving this problem by providing some kind of explanation interface to humans that have to interact with AI systems. Although this has proven to be a good idea for some purposes and has generated useful tools, such as layer-per-layer visualization of the patterns that are being recognized in deep convolutional networks, there is also a potential risk associated to the creation of a complex intermediate explanation layer, since it is difficult to measure how our decisions are impacted by it. Human decision-making is in fact known to be affected by many cognitive bias, for instance the so-called *framing effect*, which happens when our decisions are influenced by the way information is presented. If explanation interfaces are not carefully crafted they could, in principle, leverage this kind of cognitive bias to convince a human operator that the internal model is correct, or find a way of presenting it so that the predictions made by the AI are not questioned by human operators.

I. INTRODUCTION

AI has made giant steps since its birth in the late 1950s, 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” 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 errors.

One well known issue of modern AI-based decision systems is that they are difficult to debug, partly because 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 it to be able to undertake in the future.

Explainable Artificial Intelligence (XAI) is a 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 and comprehensive definition of interpretability, which is a quite discussed topic in the field. This lack of a common definition has led many researchers to concentrate only on partial aspects such as the explanation's *complexity*, which is a relatively easy aspect to measure when using proxies such as the number of elements that compose the explanation.

More difficult yet essential aspects to evaluate should be the explanation's *fidelity*, i.e. the adherence of the explanation to the original system, and its *informativeness*, which is the ability to transfer to the user a faithful representation of the internal model. But since fidelity and complexity are often in contrast when devising explanation systems, it is easy to image that the generation of a simple explanation for a complex decision can lead to omitting some aspects, that may or may not be important for evaluating it.

Taking this approach to its extreme consequences, I want to argue that a system where fidelity is never taken into account and the only metric is user feedback on the clarity of the explanation 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. This can be especially harmful for “thick” explanation layers that try to bridge the gap between human thinking and machine learning systems, which are generally very different. If these explanation systems are able to leverage human cognitive biases, which are known to exist in this kind of situations, to sound more convincing, we could end with machines that are actually incentivized to lie.

The final goal of this paper is to offer a reflection on how complex the evaluation process for XAI systems is, and how dangerous are the consequences of underestimating the biases that an explanation can introduce in the decision process in which they are involved.

The reminder of this paper is organized as follows.

Section II provides some background on the kind of problems that XAI tries to solve and a classification of the solutions that are currently being developed.

Section III introduces the problem of defining interpretability, and proposes a classification of the aspects that define an explanation.

Section IV discusses the idea that explanation interfaces might be able to fool a human user into believing that a specific algorithm is doing the right thing.

Finally, Section V contains a list of possible critiques to the ideas expressed in this paper and proposes possible improvements.

II. BACKGROUND

A. The need for interpretability in AI

Modern complex AI techniques, such as deep learning and genetic algorithms, are naturally opaque, yet they constitute the foundations of today's image and speech recognition, natural language processing, autonomous driving and many other intelligent systems. These are the kind of algorithms which are being considered in this paper.

As this technology advances, we are starting to see the problems that derive from its opacity. In the 2010s public concerns about racial and other bias in the use of AI for criminal sentencing decisions and findings of creditworthiness have led to increased demand for transparent artificial intelligence. As a result, many academics and organizations are developing tools to help detect bias in their systems.

Marvin Minsky et al. raised the issue that AI can function as a form of surveillance, with the biases inherent in surveillance, suggesting HI (Humanistic Intelligence) as a way to create a more fair and balanced "human-in-the-loop" AI.

From a practical point of view, modern AI has been found to be particularly vulnerable to "Clever Hans" situations, in which an AI chooses the right answer for the wrong reason. A notorious example is Lapuschkin et al. [4], where the Fisher vector classifier trained on the PASCAL VOC 2007 data had learned to recognize horses by recognizing a copyright tag that was present in about one-fifth of the horse figures in the training dataset.

It is important to notice that the problems highlighted in the above examples cannot be associated to any programming error of these systems, but is more of an inherent issues of these machines.

B. The XAI approach

The term *Explainable AI* refers to methods and techniques in the application of artificial intelligence that aim at improving the possibility for humans to understand its solutions. We can consider Gunning [2] as a starting point for modern XAI research. Figure 1 reports the goals of XAI as stated in that paper, which can be summarized as:

- *Accountability*: the possibility to identify precise responsibilities when things go wrong in AI.

- *Transparency*: the possibility to understand the reasons why AIs behave in a certain way, for both debugging and trust purposes.
- *Fairness*: the possibility to measure possible bias that has been picked up by the AI model.

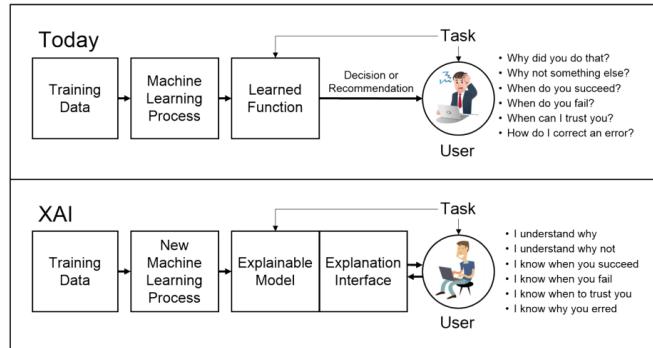


Fig. 1. The goals of XAI as expressed in Gunning [2].

The creation of such systems requires the joint effort of 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.

C. Proposed Solutions

Given the variety of goals of in this field, many different solutions have been proposed by various papers in the framework of XAI, which vary greatly in intended use and adopted approach. Guidotti et al. [1] proposes the following high-level classification for XAI techniques, as shown in Figure 2:

- *Model explanation*: Global explanation of the whole model's behavior.
- *Outcome explanation*: Local explanation of a single outcome.
- *Model inspection*: Explanation through introspection of the model's internals.
- *Transparent Box design*: Making AI easier to understand by design, e.g. by using algorithms that are intrinsically clearer for humans.

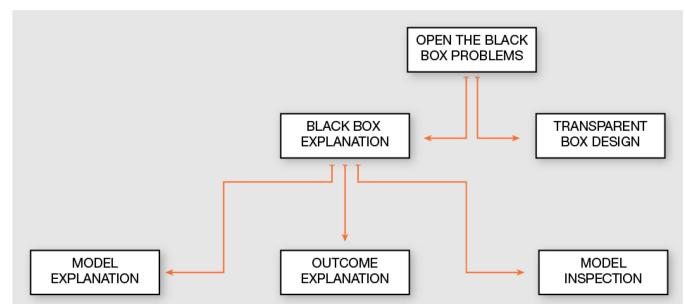


Fig. 2. Taxonomy of XAI solutions provided by Guidotti et al. [1]

This paper focuses in particular on those methods that are based on reverse-engineering already existing solutions, also

called *post-hoc interpretability*, to create an explanation. This kind of systems are the ones that are being mostly developed in the current XAI framework (Guidotti et al. [1]).

Existing solutions of this type might be classified referring to Gunning [3]:

- 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 3, to highlight which are the most significant portions of the input for the AI model.

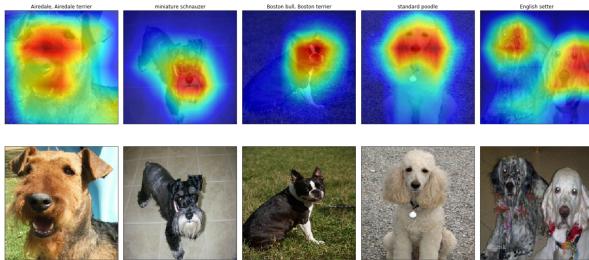


Fig. 3. An example of a heatmap representing the importance of pixels for a given classification decision, from blue (low importance) to red (high importance).

- 2) *Approximation*: this approach consists in using simple models for the explanation, or simplifying already existing models to just a set of important features. This is the case of *single tree approximation*, where the internal structure of an AI algorithm is approximated to a single classification tree, such as in Figure 4.

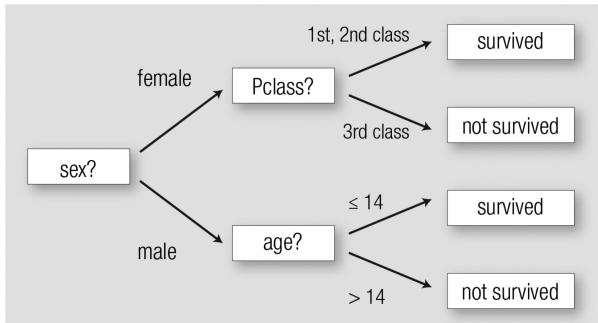


Fig. 4. A simple Decision Tree

- cit**
- 3) *Causal Models (CAMEL)*: the idea of this approach is to generate causal explanations of ML operation and present them to the user as intuitive narratives in an interactive, easy-to-use interface grounded in cognitive engineering theories. A scheme of the architecture needed for this approach is illustrated in Figure 5.
 - 4) *Learning and Communicating Explainable Representations*: explanations themselves are learned as a separate part of the training process.
 - 5) *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*,

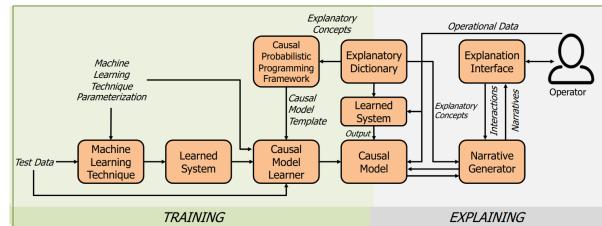


Fig. 5. Architecture of the CAMEL approach.

of how the AI thinks that a typical member of a given class should appear and/or which characteristics should be changed to change the outcome of the classifier.

III. DEFINING EXPLAINABILITY

As anticipated in Section I, one fundamental problem in the field of XAI is that there is no single conventional notion of explainability. Lipton [5] goes as far as considering the term itself ill-defined, therefore stating that claims about interpretability generally have a quasi-scientific nature. Guidotti et al. [1] on the other hand, considers the lack of a mathematical description as an obstacle for the future development of this field. Gunning [2] 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.

When analyzing the problem of defining and evaluating interpretability, two questions naturally arise:

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
- 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 and backtracking them to a specific reason
- *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.

A. 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 it, 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 influences the kind of system that we are going to build.

IV. THE TROUBLES OF EXPLANATIONS

A. Measuring fidelity

- 1) Reality
- 2) Human model
- 3) Machine model
- 4) machine explanation

B. Decision making biases

- 1) Framing effect
- 2) Post-hoc interpretations
- 3) Believe more to complex explanations you don't understand

C. How can an explanation leverage bias?

V. 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.

REFERENCES

- [1] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL <https://doi.org/10.1145/3236009>.
- [2] David Gunning. Darpa's explainable artificial intelligence (xai) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3308446. URL <https://doi.org/10.1145/3301275.3308446>.
- [3] David Gunning. Xai for nasa, 2018. URL <https://asd.gsfc.nasa.gov/conferences/ai/program>
- [4] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL <http://www.nature.com/articles/s41467-019-08987-4>
- [5] Zachary Lipton. The mythos of model interpretability. *Communications of the ACM*, 61, 10 2016. doi: 10.1145/3233231.
- [6] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 12 2017. ISSN 2044-3994. doi: 10.1093/idpl/ixp022. URL <https://doi.org/10.1093/idpl/ixp022>.

¹This goal is not explicitly listed in the original scope of XAI, but has gained traction recently with the introduction of the concept of *right for an explanation* in Europe's new GDPR. (Selbst and Powles [6])