Explainable AI and cognitive bias: why should we be careful in trusting AI explanations

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

***Abstract*—Artificial Intelligence is known to suffer from explainability 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.**

I. Introduction

add comment on interpretability vs explainability

add comment on the fact that internals are useful only to developers

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. cit This increase in AI’s capabilities is a game-changer for technology and society, so much that expressions like “AI singularity” cit 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 cit 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 cit, 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 cit 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 incentivezed 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 cit 24 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 cit 1 to increased demand for transparent artificial intelligence. As a result, many academics and organizations are developing tools to help detect bias in their systems.

cit 22 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.

cit 23 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. [5], 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.

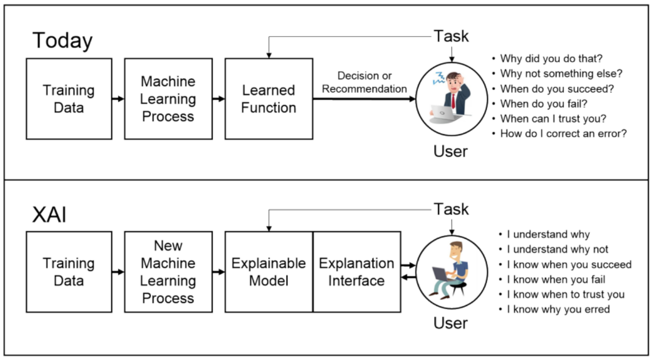

Fig. 1. The goals of XAI as expressed in Gunning [2].

The creation of such systems requires the join 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.

This paper focuses is in particular on those methods that are based on reverse-engineering already existing models, 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 to visually represent key aspects of the AI behavior. One example is to use *saliency masks*, such as the one in Figure 2, to highlight which are the most significant portions of the input for the AI model.
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 cit structure of an AI algorithm is approximated to a singe classification tree, such as in Figure 3.
3) *Causal Models (CAMEL)*: the idea of this approach cit 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 cogni-

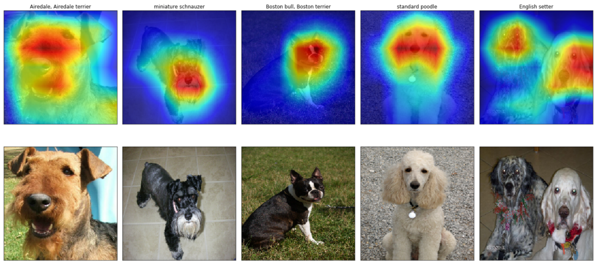

Fig. 2. An example of a heatmap representing the importance of pixels for a given classification decision, from blue (low importance) to red (high importance).

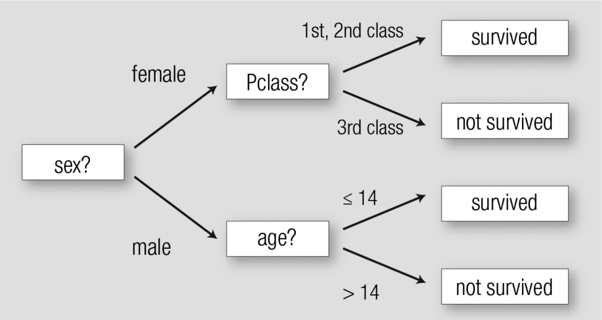

Fig. 3. A simple Decision Tree

tive engineering theories. A scheme of the architecture needed for this approach is illustrated in Figure 4.

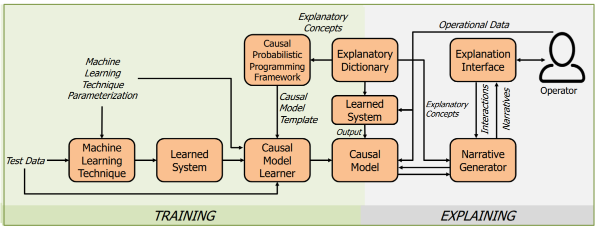

Fig. 4. Architecture of the CAMEL approach.

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*, 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.

While solutions of type 1 (Visualization) give a clear insight on what the AI model is actually doing, they are really effective only in image-based applications. Other types of explanations are more general, but more problematic. In particular, solutions of type 2 (Approximation) suffer from the possibility of over-simplification, which could be misleading, and don't directly represent the internals of the AI model. Solutions of type 3 (Causal Models) are even more distant from the original model, as they try to bridge human-like causal reasoning and statistical correlation, which is the way in which machine learning algorithms generally work. Solutions of type 4 (Learning) add another dimension to this problem, since there is the possibility for machines to *learn* the explanation representation.

These examples already show how the problem of interpretability has multiple dimensions, and the closer we try to get to human reasoning and focus on the user's thinking process, the less we are being explicit on the AI's internals and what is really happening inside the black-box.

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 [6] 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 task, 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

[1] This 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 [7])

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 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, each metric can be more or less important for specific use-cases, but there is also a difference in the way we can measure these characteristics.

Complexity, for instance, is often measured using a proxy such as the number of elements in the explanation, which can be for example the depth of the decision tree in those explanation interfaces that make use them. On the other hand, clearness and informativeness are more difficult to quantify a-priori, but could be empirically evaluated by providing explanations to a human control-group and verify how they respond.

In general, we can identify two ways of evaluating an AI explanation: one is using a direct measurement of some quantity that we can derive directly from the explanation. The second one is considering an explanation itself a black-box, and check if it actually provides a better understanding of the AI model to some selected group of individuals used as a benchmark. While the first method is not always feasible, since choosing which quantity is representative of a certain aspect is in itself a difficult decision to make, the second method clearly presents the same problems of opaqueness and unreliability that AI models themselves have.

IV. The Troubles of Explanations

A. Measuring fidelity

Of all the metrics highlighted in Section III-A, fidelity, also called *faithfulness* in literature , is probably the most complex aspect to evaluate. On one hand, the maximum fidelity is already represented by the implementation itself, but on the other the reason we need explanations is that the implementation itself is not clear enough. Additionally, it is difficult to isolate this aspect in explanations, as it has to do both with the explanation interface and with how the human user interprets the explanation.

Yet, fidelity plays a fundamental role when we have to evaluate an AI algorithm, as it quantifies the difference between what is being evaluated (the AI model) and the instrument we are using for this evaluation (the AI explanation). Figure 5 attempts to represent the aspects of AI evaluation through a explanation interface: in this case, a human evaluator needs to compare the AI explanation with his own model of the world and, based on this information, evaluate the correctness of the AI model.

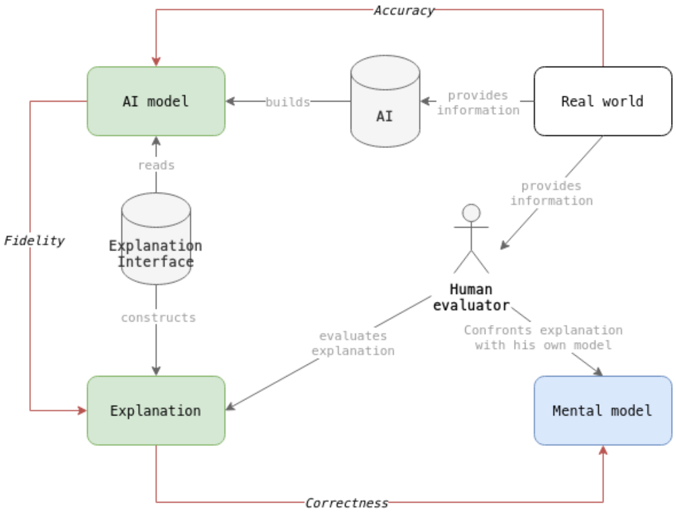

Fig. 5. Evaluation process of an AI through an explanation interfaces.

A problem that can be present in this situation is the fact that the AI model and the representation chosen for the explanation could be of very different nature, such as in Causal Models described in Section II-C. If that is the case, does it even make sense to ask ourselves if a causal relationship “really” represents the AI model, which typically learns using statistical correlation? And if, on the other hand, both the AI and the explanation are black boxes, how can we be sure that evaluating the AI using that explanation interface will effectively improve our understanding of the underlying AI model? Could we just *think* we are understanding it?

B. Decision making biases

Human decision-making is known to be affected by many cognitive bias, which are deeply rooted in our thinking and are often difficult, if not impossible, to exclude when we make decisions. Recently, Kim and Song [4] has studied the consequences of the *framing effect* in the domain of AI, in particular how likely is a person to accept or reject an AI recommendation based on how the output was framed. An interesting result of this research is, for example, that “*perceived reasonableness was significantly higher when the suggestion of AI was provided before the decision is made* *than after the decision is made when perceived accuracy was controlled*".

While this is not a direct study on AI explanation interfaces, it does show how the same local decision of an AI can be judged differently just varying the timing of the explanation, and a similar result is found about the way the solution is framed.

This goes to show how the evaluation of the correctness of an AI model is not only a subjective matter, but can vary in the same individual depending on factors that are external to the AI behavior itself.

C. Can explanation interfaces learn to exploit cognitive bias?

As a pure thought experiment, let's imagine a situation where a single individual is in charge of deciding whether an explanation interface is suited for understanding a certain type of AI architecture, e.g. Neural Networks, and let's make the assumption that the only observable elements are the output of the AI model and the explanation provided by the interface. This settings is not far from the one of an operator who acts upon AI recommendations. Let's then suppose that the explanation interface is a complex interface that can convert the model's internals into a human-like reasoning, for example by producing textual motivations for a certain output.

As we demonstrated, it is completely possible that the judgment about the correctness of the algorithm is biased by the way the explanation interface presents the information. Let's now suppose that the same individual has to choose between multiple explanation interfaces and multiple AI models the best couple of model-explainer to be put in production: even if the single explanation interface is not built to learn from the individual's taste, this environment creates a selection for those explanation systems that present information in a way that fits better the idea the individual has about the problem. In particular, an feasible solution to this problem could be a couple in which the explainer is very convincing at justifying the AI model. In absence of other information, the individual has equal probability of selecting the most correct explanation interface and the most accurate AI model, or the most convincing explanation interface coupled with a sub-optimal model.

In this way, we have created a scenario in which an explanation interface that can lie very well is indistinguishable from a very accurate explanation of a convincing model.

V. Conclusions

In conclusion, we have showed how the fact that there is no single definition of what interpretability is and no comprehensive way of evaluating all important aspects that compose an explanation leads to the possibility of creating yet another black-box layer over the black-box model, which can accentuate biases instead of reducing them.

Some criticisms that can be made to this argument are:

- Does it mean that all XAI solutions are useless? Absolutely not. This is an important research that is fundamental for the future of AI, and the fact that many directions are being explored to solve this problem is just a reflection of how multi-faceted is the problem itself.
- Can't the problem stated in IV-C be overcome by simply adding more judges? While this is true, the problem of cognitive bias is that it can also be inherited from the group or society around the single individual. Certainly, a diverse control group for evaluating explanations is a must-have, but it is not a complete guarantee of absence of bias.
- Can't errors introduced by the explanation interface be simply corrected by looking at the actual behavior of the AI? If the explanation itself behaves more or less like a black-box, i.e. in those cases in which it is difficult to quantify the explanation's fidelity, there is the same problem that affects AI testing in general: even if testing is done extensively and all results are positive, i.e. the explanation is always perfectly aligned to the AI model, we have no general guarantee that this holds for all the possible outcomes.

In the end, while the proposed argument is just a thought experiment, there are many elements which are realistic and should be paid attention to when devising XAI solutions.

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] Taenyun Kim and Hayeon Song. The effect of message framing and timing on the acceptance of artificial intelligence's suggestion. 04 2020. doi: 10.1145/3334480.3383038.

[5] 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-

[6] Zachary Lipton. The mythos of model interpretability. *Communications of the ACM*, 61, 10 2016. doi: 10.1145/3233231.

[7] 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/ipx022. URL https://doi.org/10.1093/idpl/ipx022.