Explainable AI and cognitive bias: opportunities and perils of AI explanation

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

***Abstract*—Artificial Intelligence (AI) has become an increasingly significant part of our societies and private lives nowadays. From curation systems that can influence what we buy, what we watch, what we believe in and who we vote for, to systems that can decide who to hire, predict criminal recidivism, and operate on the stock market, many fields have adopted some kind of AI system for tasks that would normally require human intelligence. However, this kind of technology is known to suffer from interpretability problems: it is often hard for humans to understand why an AI-based system took a particular decision and to identify the general logic behind an AI's behavior. XAI (short for eXplainable AI) is a research field that aims at finding methods to explain these machines' behaviors or individual decisions in a way that is intelligible to humans. Although there are great opportunities in this field, there is also a risk associated to the creation of intermediate explanation layers. Human decision-making is in fact known to be affected by cognitive bias, for instance the so-called *framing effect*, which happens when our decisions are influenced by the way information is presented. This paper wants to explore how explanation systems could, in principle, leverage decision-making biases to convince human operators that their internal model is correct, or present decisions in a way that they are not questioned, if these explanation interfaces are not cautiously evaluated.**

I. Introduction

Artificial Intelligence is an umbrella term that is commonly used to describe a multitude of technologies and approaches. While a precise definition of this term involves many complex aspects, such as defining the meaning of intelligence and its relation with human thinking (Russell and Norvig 2009, chapter 1.1), in this paper I want to focus on those computing techniques that are currently used to take automated decisions in complex situations, i.e. in which the desired behavior cannot be easily synthesized using simple rules.

One of such techniques is Machine Learning, which has enabled many of the recent advances in the field. It is used to recognize handwritten digits, people and objects in pictures, translate from one language to another, drive cars, and suggest photos, videos and posts on platforms such as Facebook and YouTube.

While, on one hand, these techniques have achieved incredible results, and the surrounding technology has become mature enough to be employed in fields such as medicine, law and finance, on the other hand there is a lot of research that has highlighted the social, cultural and political implications of some of the side-effects of ML-based systems, for example when they learn biased models or spurious correlations.

Yet, most of the popular top-performing algorithms in use today, such as Deep Convolutional Neural Networks (DCNNs), Genetic Algorithms, Swarm Intelligence and Statistical Machine Learning in general, are known to be difficult to examine and understand for humans.

Most of these algorithms are in fact built to reason in a statistical way, are composed of a large number of elements (e.g. neurons in DCNNs) and can have very complex structures.

This aspect represents a huge technical and ethical issue for this field, especially when building autonomous systems that are meant to replace or aid humans in highly impacting decisions. If we can't explain "why" a certain algorithm took a certain decision, how can we trust these systems? How do we ensure that their internal models are not biased or broken? How do we understand when the machine is failing?

The problem of introspection and accountability for these systems is a very serious one. 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 (Minsky, Kurzweil, and Mann 2013).

As a natural result of these emerging concerns about AI, the field of Explainable AI (XAI) was born. The goal of this research field is to build systems that can provide humans with a deeper understanding of AI algorithms, with the ultimate objective of making errors and biases more easy to spot or predict and AI-based systems generally more trustworthy.

Yet, because of the young age of this field and because of the intrinsic complexity of the problem, there is still a lack of definition of what *explainable* means, and there is a somewhat confused terminology around the idea of *interpretability*.

In this paper, I want to analyze some of the extreme consequences of the lack of such definition, and more generally the lack of a comprehensive way to evaluate AI explanations.

Since there is no widely accepted way of evaluating explanations, each researcher tends to focus on one or just a few aspects of interpretability. (Guidotti et al. 2018) in particular shows that many studies concentrate on the *complexity* of the explanation. I want to show in this paper that an evaluation system based solely on how easy it is to understand an explanation, without taking into account aspects such as fidelity, might produce potentially harmful explanation interfaces.

In particular, I want to show that the presence of well known cognitive bias in the way humans understand and accept explanations can distort the evaluation of these explanations,

and if this aspect is not taken into account, AI explanation could amplify the problem of unconscious biases in technology instead of mitigating it.

To explain this idea, I will proceed as follows:

In Section II, I will provide some background on the problems which XAI is trying to solve and a classification of the solutions that are currently being developed.

In Section III, I will introduce the problem of defining interpretability, and propose a classification of the aspects that define an explanation.

In Section IV, I will discuss the idea that explanation interfaces might be able to fool a human user into believing that a specific algorithm is doing the right thing, leveraging his or her own bias.

Finally, Section V contains a list of possible critiques to the ideas expressed in this paper and Section VI contains some concluding thoughts on this subject.

II. Background

A. Current issues of AI

As AI has become more popular, there has been a vast number of studies on potential or actual negative externalities of real-world AI systems. One well-known category is curation and filtering algorithms for online platforms and search engines, for example YouTube.

(Tufekci 2018) claims that "Youtube may be one of the most powerful radicalizing instruments of the 21st century", and automated suggestions play a key role in this, since 70% of the videos watched are recommended by an ML-based system (Solsman 2018).

(Epstein and Robertson 2015) found that biased search engine results can shift the voting preference of undecided voters by 20%.

Challenges such as fake news, biased predictions and filter bubbles make the understanding of ML-base curation systems an important and timely concern, but there are even more sensitive contexts where defects in AI systems can provoke disastrous consequences.

Some recent studies have shown how subtly gender, race and social biases can be inherited by ML algorithms: facial recognition is known to have biases with respect to skin color (Buolamwini and Gebru 2018), Amazon's hiring algorithm was shown to disproportionally prefer men to women (Dastin 2018), and the COMPAS algorithm used to estimate criminal recidivism has been accused of racial bias (Julia Angwin and Kirchner 2018).

More generally, we know that current AI is prone to what are called "Clever Hans moments": a notorious example is (Lapuschkin et al. 2019), where the classifier taken into consideration had learned to recognize image of horses because there was a copyright tag that was present in about one-fifth of the horse figures in the training dataset.

This goes to show that, despite the advanced results that have been achieved with AI and Machine Learning, there are still many problems that require the creation of introspection and explanation tools in this field.

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 AI algorithms. We can consider (Gunning 2017) as a starting point for modern XAI research. Figure 1 reports the goals of XAI as stated in that paper. The idea is that we want to be able to better understand when a given AI-based system is doing something wrong, when we can trust it and why an error occurred. All these aspects contribute to improving *accountability*, since who makes use of such systems can and must verify their behavior, *transparency*, since the reasons for a certain decisions are in theory stated in an understandable way, and *fairness*, since understanding the reasons enables us to oppose a certain decision.

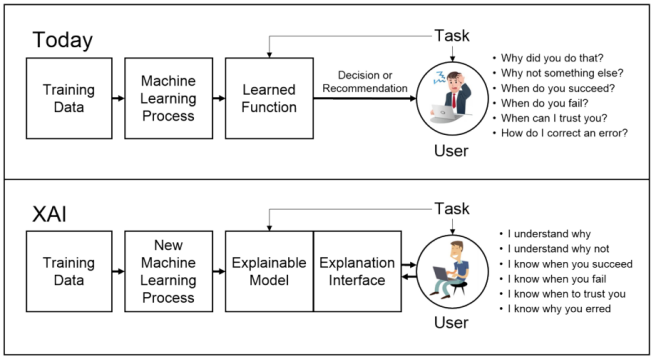

Fig. 1. The goals of XAI as expressed in (Gunning 2017).

C. Proposed Solutions

From a general perspective, (Guidotti et al. 2018) identifies two families of approaches to this problem: *transparent box design*, which aims at building algorithms that are more explainable by design, and *reverse-engineering* approaches, also called *post-hoc interpretability* approaches, which try to provide explanations for already existing algorithms.

Some examples of the latter type are listed in (Gunning 2018).

Visualization, for instance, focuses on representing visually some key aspects of the model, for example which pixels of an image are important for a classification output, as shown in Figure 2.

Approximation consists in using simple models or simplifying already existing models: in *single tree approximation*, for example, the internal structure of an AI algorithm is approximated to a classification tree, shown in Figure 3.

Causal Models (CAMEL) try to generate causal explanations of Machine Learning operations and present them to the user as intuitive narratives.

Other approaches include *Learning and Communicating Explainable Representations*, where explanations themselves are learned as a separate part of the training process, and *Explanation by Example*, where the AI is able to provide an example, or a *prototype*, of how it thinks that a typical member

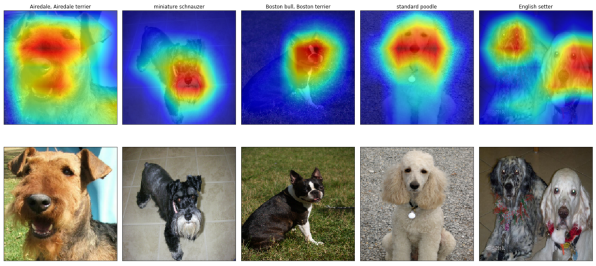

Fig. 2. An example of a heatmap representing the importance of each pixel in determining if a picture has been classified as containing a dog, from blue (low importance) to red (high importance).

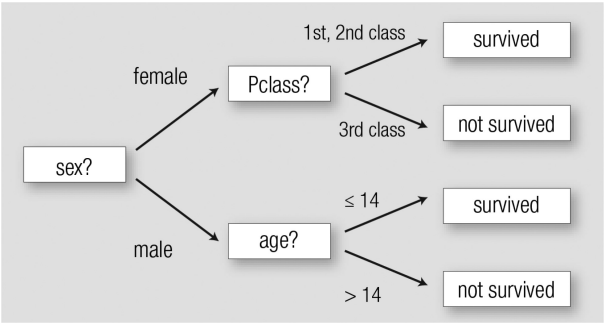

Fig. 3. A simple Decision Tree

of a given class should appear and/or which characteristics should be changed to change the outcome.

It is important to notice how these approaches differ in how *thick* the explanation interface is, i.e. how many complex manipulations the initial model undergoes before being presented to the user. Intuitively, we can see for example that the visualization approach tries to give a close insight on how the internal elements are activated by a certain picture, while in techniques such as CAMEL and Learned Explanations there is a much more indirect connection between elements of the original model and elements of the explanation, which is also reflected on the increased complexity of the interface itself.

This intuitive idea will be further explored in Section IV-A with the concept of *fidelity*.

III. Defining Interpretability

A. Dimensions of AI Explanation

As anticipated in Section I, one fundamental problem in the field of XAI is that there is no single conventional notion of interpretability. (Lipton 2016) 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. 2018) on the other hand, considers the lack of a mathematical description as an obstacle for the future development of this field. (Gunning 2017) itself defines the formalization of an evaluation metric for explanations as one of the goals of the XAI program, 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, which may partially overlap. Therefore there is a variety of goals that explanations try to accomplish, and some of them might be in contrast with each other. Some of these needs 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
- *Failure Prediction*: understanding which are the weaknesses of an AI system and when it is likely to fail
- *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 built with different users in mind will have very different notions of what a good explanation is.

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 functioning of the system? Are all the facts inferred from the explanation also applicable to the original system?

Clearly, a specific metric will be more or less important depending on the specific user and use-case. There is however a deeper distinction that has to be made, which is related to how these metrics are measured.

[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 2017)

Complexity, for instance, is often measured using a proxy quantity such as the number of elements in the explanation, which can be for example the depth of the decision tree or the number of neurons. On the other hand, clearness and informativeness are more difficult to quantify a-priori, but could be empirically evaluated by providing the explanations to a group of humans and verifying 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-B, fidelity, also called *faithfulness* in literature (Gilpin et al. 2018) , is probably the most complex to evaluate. On one hand, the maximum fidelity is already represented by the implementation itself, but on the other hand the reason we need explanations is that the implementation itself is not clear enough.

This is particularly important since AI explanations are also targeted to unspecialized users, which need to understand what's happening without necessarily having a solid background on the internal functioning of such systems.

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). This is what we intuitively defined as *distance* between the model and the explanation in Section II-C, and represents in some sense the "measurement error" introduced by the explanation.

While this idea might seem easy enough to understand, devising an operational way to measure it is a non-trivial task.

Let's take for example Causal Models: in this case, the explanation and the original model will typically have a very different nature, since the explanation interface produces causal relationships, while the AI model typically reasons in terms of statistical correlation. In this case, how can we measure the fidelity of this interface?

On the other hand, being unable to measure fidelity poses another question: if both the AI and the explanation are treated as 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? Couldn't it be that we just *think* we understand 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 2020 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*" (Kim and Song 2020, page 5).

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 has been found when looking to how the solution is framed (positive or negative sentences etc.).

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 setting 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 shown in Section IV-B, 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 select, 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 is perceived better by the operator.

Although in the ideal case this leads to selecting an explainer that is clear and comprehensible for a human, a possible outcome could also be selecting 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.

This scenario effectively describes a situation in which we have created an explanation machine which is so good at producing convincing explanations that it could also *lie* without being detected.

V. Counterarguments

Some criticisms that can be made to this argument are:

- *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 interface behaves like a black-box, 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.
- *Can't we just use explanations that have a high fidelity with respect to the AI model?* While this is a possible solution when the explanation is used for debugging purposes, we should not forget that the audience of AI explanation is much broader, and it is especially important that non-expert people that are subjected to the decisions of an AI system are given the possibility to understand them.

VI. Conclusions

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

While the proposed argument is just a thought experiment, there are many realistic elements in this setting that should warn us about the possibility of creating deceitful explanation interfaces.

References

[1] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91.

[2] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (Oct. 2018). URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[3] Robert Epstein and Ronald E. Robertson. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections". In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (Aug. 2015). DOI: 10.1073/pnas.1419828112.

[4] L. H. Gilpin et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018, pp. 80–89. DOI: 10.1109/DSAA.2018.00018.

[5] Riccardo Guidotti et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Comput. Surv.* 51.5 (Aug. 2018). ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009.

[6] David Gunning. "DARPA's Explainable Artificial Intelligence (XAI) Program". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2017, p. ii. ISBN: 9781450362726. DOI: 10.1145/3301275.3308446. URL: https://doi.org/10.1145/3301275.3308446.

[7] David Gunning. *XAI for NASA*. 2018. URL: https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf.

[8] Surya Mattu Julia Angwin Jeff Larson and Lauren Kirchner. "Machine Bias". In: *ProPublica* (2018).

[9] Taenyun Kim and Hayeon Song. "The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion". In: Apr. 2020. DOI: 10.1145/3334480.3383038.

[10] Sebastian Lapuschkin et al. "Unmasking Clever Hans predictors and assessing what machines really learn". en. In: *Nature Communications* 10.1 (Dec. 2019), p. 1096. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08987-4. URL: http://www.nature.com/articles/s41467-019-08987-4 (visited on 12/15/2020).

[11] Zachary Lipton. "The Mythos of Model Interpretability". In: *Communications of the ACM* 61 (Oct. 2016). DOI: 10.1145/3233231.

[12] M. Minsky, R. Kurzweil, and S. Mann. "The society of intelligent veillance". In: *2013 IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life*. 2013, pp. 13–17. DOI: 10.1109/ISTAS.2013.6613095.

[13] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. USA: Prentice Hall Press, 2009. ISBN: 0136042597.

[14] Andrew D Selbst and Julia Powles. "Meaningful information and the right to explanation". In: *International Data Privacy Law* 7.4 (Dec. 2017), pp. 233–242. ISSN: 2044-3994. DOI: 10.1093/idpl/ipx022. eprint: https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf. URL: https://doi.org/10.1093/idpl/ipx022.

[15] Joan Solsman. "YouTube's AI is the puppet master over most of what you watch". In: *cnet* (2018).

[16] Zeynep Tufekci. "YouTube, the great radicalizer". In: *The New York Times* 10 (2018), p. 2018.