A Philosophical Approach for a Human-centered Explainable AI[1]

Luca Capone[1] and Marta Bertolaso[2]

[1] University Campus Bio-Medico of Rome, RM, 00128, Italy.
[2] University Campus Bio-Medico of Rome, RM, 00128, Italy.
l.capone@unicampus.it
m.bertolaso@unicampus.it

Abstract. Requests for technical specifications on the notion of explainability of AI are urgent, although the definitions proposed are sometimes confusing. It is clear from the available literature that it is not easy to provide explicit, discrete and general criteria according to which an algorithm can be considered explainable, especially regarding the issue of trust in the human-machine relationship. The question of black boxes has turned out to be less obvious than we initially thought.
In this position paper, we will propose a critical analysis of two approaches to Explainable AI, a technically-oriented one and a human centered model. The aim is to highlight the epistemological gaps underlying these proposals. Through a philosophical approach, a new starting point for Explainable AI related studies will be handed out, which will eventually be able to hold together the technical limits set by algorithms and the instances of a human-centric approach.

Keywords: XAI, Black-Box, Philosophy of Technology.

1 Introduction

In the last thirty years, digital technologies have held a leading position in the race to automation, where artificial intelligence and machine learning algorithms are now taking over. These tools have been used in the most disparate fields, in scientific research [2], hoping to be able to discover causal links starting from correlations; in logistics and in the administration of companies [11], up to more everyday scenarios that impact increasingly larger segments of the population. We find algorithms employed in the medical, jurisprudential and economic fields [16], all these uses have raised concerns about the reliability of the systems in their relationship with users.

The problems unfolded by the uses of AI in these fields are of a pragmatic kind, about the correct functioning of these technologies and the possibility of detecting potential errors, of a legal nature, regarding the responsibility of decision-makers who

rely on these systems, and of an ethical one, concerning the possible discrimination that these technologies could produce. In the wake of these questions, a research strand dedicated to explainable AI (XAI) has been developed, which deals with methods and procedures aimed at explaining, examining and justifying the work of artificial intelligence systems.

In this position paper, although adhering to Miller's minimum XAI definition of "explanatory agent revealing underlying causes to its or another agent's decision making" [12], we will depart from his empirical approach and will analyze the theoretical assumptions that have informed the various ramifications of this field of studies.

Throughout the available literature on the subject, two different approaches to the problem can be immediately identified, corresponding to the two sides of the issue. The first one, which could be defined as the *technical* approach, focuses mainly on algorithms and the requirements which allow them to be defined as explainable (interpretable, transparent). The second one, on the other hand, the *human-based* approach, takes on the question of the human-machine relationship. Although in both approaches there is some attention to the clarification of the term *explaination*, the human centered one focuses more on the way users approach systems, their way of rationalizing outputs and interpreting the information provided by machines. Both methods, in their partiality, highlight important sides of the problem, but more interestingly, they are united by a common demand for terminological clarity. According to these studies, requests for technical specifications on the notion of explainability of AI are urgent, although the definitions proposed are sometimes confusing.

The methodologies adopted are very heterogeneous, there are those who try to outline a taxonomy for the terms under examination [1], some try to formulate a technical definition related to precise requirements that the algorithms should meet [14], others propose an explaination model based on conversational [15] or cognitive [3] structures. It is clear from the scientific literature that it is not easy to provide explicit, discrete and general criteria according to which an algorithm can be considered "explainable", especially with regards to the issue of trust in the human-machine relationship [7]. The question of black boxes has turned out to be less obvious than we initially thought.

This work's focus will be the highlighting of the conceptual biases underlying the two aforementioned approaches through a critical and philosophical analysis. Specifically, two biases underlying the methods of explaination examined will be analyzed. On the one hand there is the illusion that a term belonging to ordinary language might have a unique technical definition, a regularity of use which, however, does not belong to it. This prejudice is closely linked to a second mentalist misconstruction, the idea that behind our common speech, behind the terms we use to describe our behaviors, mental processes are hidden, in a one-to-one relationship with our behavior. Therefore, coherently to this, in order to explain an algorithm one should look into it. Interestingly, this is the same conceptual error behind the GOFAI (good-old-fashioned artificial intelligence) [4]. To clarify these prejudices, these dominant metaphors [6] within the XAI debate, will serve to clear the ground for future research on the subject.

2 A Technical Definition for XAI

It is possible to imagine the strand of works in XAI as an imaginary line at the extremes of which there are, on the one hand, the human-centered approach and on the other the technical approach. In this section we will deal with the latter.

Although this is a mere descriptive expedient, and there are no studies that completely ignore the opposite side of their point of view, we can still exemplify the technical approach to the problem of explainable AI based on some characteristic features:

- The concern for a unique and shared definition of terms (explaination, interpretation, transparency).
- The subdivision of terms into taxonomies that show precise similarities and differences.
- The concern about the clarification of precise technical characteristics that may be valid as requirements to declare a model as transparent, an algorithm as explainable, a decision-making assistance system as interpretable.

These three factors are obviously interlinked and try to answer to the problem of interpretability according to a coherent reasoning, which we could reformulate in this way: to address the problem of the intelligibility of the work of algorithms it is necessary, first of all, to define what we mean by the term explaination. Once we have defined what we mean by explaination and specifically, explicability of an algorithm, we will have to list the characteristics that it must have, or the conditions under which an algorithm can be defined explainable, or even the procedures by which it can be made so. Although at first glance it may seem a reasonable program, the literature has already raised several critical issues in this regard, which we list below.

The heterogeneous composition of the users population makes the suggested explainations excessively standardized. Specifically, the risk is that technicians are designing explainations that work for them, but not for those who will have to use these technologies and even less for those who will simply be subject to them [13]. A further criticism is that this kind of solutions would lose sight of the socially situated nature of these tools, providing models of explainations designed for ideal situations [7], with the risk of losing contact with the real world.

Although it might be useful to go into detail on each of the previous criticisms, and many others not reported here, the current goal of this article is to critically analyze this approach, highlighting its epistemological gaps. The objection, from this point of view, is that this approach is based on an incorrect terminological assumption about the status of the term explaination. It is necessary to reflexively explore this concept and its place within the language to clarify this misunderstanding. Among the various definitions proposed in the examined papers, formulated from scratch or cited by authoritative dictionaries [14], it is possible to see how the term explaination is almost always referred to a description [3], it has a discursive character [15] and, in the end, could be described as the exposition of a series of information [9]. The problem here is of a linguistic nature.

In his Philosophical Investigations, Ludwig Wittgenstein, a famous philosopher of language, devotes several pages to the question of explaination and its pragmatic and conversational character [17].

"§87. […] One might say: an explanation serves to remove or to avert a misunderstanding – one, that is, that would occur but for the explanation; not every one that I can imagine." And shortly after, "§109. […] We must do away with all *explanation*, and description alone must take its place. And this description gets its light, that is to say, its purpose, from the philosophical problems. There are, of course, not empirical problems; they are solved rather, by looking into the working of our language, and that in such a way as to make us recognize those workings: *in despite of* an urge to misunderstand them." [17].

What Wittgenstein proposes is not a hymn to relativism, but an honest acknowledgement of the descriptive and situational nature of what constitutes an explaination. It is possible to imagine that in certain circumstances, having to explain why some products are suggested to me instead of others, it will be enough to indicate some information about my past purchases. While, to understand why I was refused a loan, it will require much more data, and sometimes more or less sophisticated technical knowledge. In both cases, the term explaination does not need a univocal definition, but performs its normal function despite its intrinsic vagueness, as if it was a conversation between human beings. Whether it is about interacting with another human being or with a machine, these explainations will amount to nothing more than descriptions, they are basically lists of facts that become explainations to the extent that users know what to do with them. It is unthinkable to have a theory of explaination [7] in the same way as we have a theory for physics. It is possible to have approaches to explaination depending on the problem to be clarified, the level of abstraction needed [8] and the target of the explaination.

The technical approach criticized here is, nonetheless, trying to answer to legitimate needs. On the one hand, we have the algorithms, which we could more precisely define as parametric functions, which can be represented as computational graphs in the case of neural networks. On the other hand, we have end-users who may have no idea of what a parametric function is, and how difficult it is to get an explaination out of it. The insistence of the technicians for a correct definition of the term explaination can be seen as a reasonable response to this tension. This leads to the next section and the need to consider the human counterpart of the problem.

3 The Human-centered Approach

As illustrated in the previous chapter, human centered approach to XAI criticizes the technical one for being too monolithic with regards to the problem of explaination, both on the issue of the terminological definition and on the assumptions of the human machine relationship. On the other hand, what is proposed is a more socially situated solution [6], which takes into account the way in which users approach machines and their interest in understanding the way algorithms work [1], their ratio.

In short, if a definition of a general explaination to be applied to algorithms was previously sought, now terminology issues are left (partially) aside, to look for the substance of this explaination directly within the algorithms, in their structure and their alleged internal mechanisms, or trying to extrapolate it post hoc, from their outputs. In the rest of the chapter how this reiterates a mentalist prejudice and how this has cost dearly in the early days of AI research will be illustrated [5].

The prejudice lies in believing that the internal operations of the algorithm can be translated into satisfactory causal explainations, or that through the outputs we can trace back to a theory of which the system would be the repository. In short, it is a matter of reducing the problem of the explaination to a mere question of convertibility. But even assuming that this is possible, is this really what an explaination amounts to? Nobody would explain his own behavior by giving a causal account of his internal states, this is not what we are looking for in an explaination. The difference here lies between a causal explaination and a justification [9].

It is worth noting two things: the first is that if we require a theory that explains the ratio according to which some phenomena can be predicted, it makes no sense to look for it within the algorithms, since they are mostly used in contexts where it is not possible to have a theory, but only a probabilistic prediction. The second is that, the type of intrinsic transparency that we would like to ascribe to these systems, has no role in the relationship between human beings, where the explainations (justifications) are always and only post hoc. Confirming indeed Wittgenstein's quotation, on the fact that every explaination can be better framed as a description. In this sense, individuals are extremely opaque.

In conclusion, even if it was possible to look into one of these systems and represent their model graphically, it is not immediately clear how the n-dimensional representation of a parametric function can help us to understand how our algorithm arrived at a certain prediction or classification. However, if we consider Miller’s idea that taking the way in which individuals provide explainations as a model might be used inductively, we have not yet come to terms with the fact that providing a post hoc explaination of an algorithm, interpreting the ratio of its predictions, is a hermeneutical exercise that still needs legitimization.

4 Conclusions

The term explaination, or rather the practice of explaining, is one of what the abovementioned philosopher calls language games. These are original human practices, which cannot be further reduced and are inextricable from the cultural, social, linguistic and pragmatic fabric of which they are made up. If we look for a definition of explaination, we must look at the way this word is used within the language, where it has its home. Keeping these ordinary practices in mind, it will then be possible, in accordance with the reference context, to formulate specific solutions to explain the algorithms' activities. There will be cases in which to make an algorithm actually transparent will be possible, and the parameters taken under consideration by the system will be available for observation. In other instances, this will not be possible, and

we should rely on a post hoc explaination able to convince us of the algorithm's work. But we should always recognize the fact that in both cases we are dealing with descriptions, whether they are textual or graphic, in natural language or in code. Once these premises have been taken into consideration, it becomes possible to formulate specific explaination procedures and justification criteria, related to specific sectors. Similarly, professional figures can be imagined, experts in a specific field, who can provide post hoc justifications if necessary. As previously seen, there are no general solutions, able to satisfy every need. Technologies will have to adjust to the procedures and contingencies of the domain in which they operate, trying to support the decision, without replacing it or hindering it.

References

1. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115 (2020).
2. Bertolaso, M., Sterpetti, F. (eds.): A critical reflection on automated science. Will science remain human? Springer (2020).
3. Doran, D., Schulz, S., Besold, T.R.: What does explainable Ai really mean? A new conceptualization of perspectives. Computer Science and Engineering Faculty Publications. Wright State University (2017). https://corescholar.libraries.wright.edu/cse/518/.
4. Dreyfus, H.L.: What computers still can't do: A critique of artificial reason. MIT Press, Cambridge (1992).
5. Dreyfus, H.L.: Why Heideggerian Ai failed and how fixing it would require making it more Heideggerian. Artificial Intelligence 171, 1137–1160 (2007).
6. Ehsan, U., Riedl, M.O.: Human-centered explainable AI: Towards a reflective sociotechnical approach. Proceedings of HCI International 2020: 22nd International Conference On Human-Computer Interaction (2020).
7. Emmert-Streib, F., Yli-Harja, O., Dehmer, M.: Explainable artificial intelligence and machine learning: A reality rooted perspective. WIREs. Data Mining and Knowledge Discovery 10(3): e1368 (2020).
8. Floridi, L.: Marketing as control of human interfaces and its political exploitation. Philosophy & Technology 32, 379–388 (2019).
9. Krishnan, M.: Against interpretability: A critical examination of the interpretability problem in machine learning. Philosophy & Technology 33, 487–502 (2019).
10. Lipton, Z.C.: The Mythos of Model Interpretability. 2016 ICML Workshop on Huma interpretability in Machine Learning, (2016).
11. Mayer-Schönberger, V., Ramge, T.: Reinventing capitalism in the age of big data. Basic Books, Ney York (2018).
12. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38 (2019).
13. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: Beware of inmates running the asylum. Or: How I learnt to stop worrying and love the social and behavioural sciences. IJCAI 2017 Workshop on Explainable Artificial Intelligence (2017).

14. Puiutta, E., Veith, E.M.S.P.: Explainable reinforcement learning: A survey. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) International Cross-Domain Conference for Machine Learning and Knowledge Extraction, CD-MAKE 2020, vol. 12279, pp. 77–95. Springer, Cham (2020).
15. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In: Trattner, C., Parra, D., Riche, N., (eds.) ACMIUI-WS 2019, vol. 2327. http://hdl.handle.net/10609/99643.
16. van Dijck, J., Poell, T., de Waal, M.: The platform society. Public values in a connective world. Oxford University Press, NY (2018).
17. Wittgenstein, L.: Philosophical Investigations. Blackwell Publishers, Oxford (1999).