ARTIFICIAL INTELLIGENCE AND THE WHY STAGE

The recent remarkable performance of Artificial Intelligence (AI), supported by Deep Learning techniques, has created expectations about its positive transforming potential in society. However, ethical and moral issues have also reappeared with great intensity. Interpretable AI emerges as a partial response to these concerns and the necessary continuous improvement.

JAIME S. CARDOSO[1,2],
LUÍS F. TEIXEIRA[1,2]
[1] INESC TEC;
[2] FACULTY OF ENGINEERING OF THE UNIVERSITY OF PORTO
jaime.cardoso@inesctec.pt
luis.f.teixeira@inesctec.pt

The so-called "why stage" is a classic and usual period in children's development. The child, eager to know the world around him, begins to question the adult about everything he wants to understand; the adult, with patience and respect, helps her to clarify her doubts, thus contributing to her learning process.
It will be exciting when Artificial Intelligence (AI) plays this role of the adult and us, the child who wants to learn. When artificial intelligence is sufficiently developed, it can, by explaining its decisions, contribute to our intellectual growth.
Until then, there is still a long way to go. AI still fails. Therefore, much of the current work in interpreting the automatic decision aims to understand the error to improve the decision algorithm and increase our trust in the machine. It is interesting to note that, in some 'closed' domains, despite deciding globally well, very well, the machine makes 'childish' mistakes and is easily manipulated. This statistically positive behaviour but with individual aberrant cases raises doubts about the concepts that the algorithm integrated; these are doubts that must be dispelled and overcome.
Much of the current work in AI deals with the so-called Deep Learning algorithms. Deep Learning is a specific area of Machine Learning, where learning algorithms generate models from the patterns found in the examples that are processed. One of the most evident differences in Deep Learning is that, in addition to learning decision models, data representation models are also learned. In other words, a model is learned that transforms the input data, for example, an image, into an abstract representation of concepts representative of that image. The performance achieved by these algorithms is remarkable, being state of the art in several domains, for example, in medical image analysis, challenging specialists in their areas. However, there is an obstacle in the interpretation of the decision process of these models - their opacity. In other words, its high complexity and high abstraction make the automatic decision challenging to interpret by humans, whether these are specialists or laypeople in medicine or AI.
In an attempt to overcome this difficulty, interpretable AI tries to justify a decision based on supplementary information. For example, it can highlight the most relevant regions of the image for decision making. The 'interpretation' algorithms provide a map of relevance (represented by an image) where the zones that conditioned the decision are identified. The calculation of these maps can be done in different ways, but always trying to assign the responsibility of the decision to the various inputs of the model. For example, suppose the model predicted cancer with an 80% confidence based on a mammogram. In that case, the interpretation models will try to assign responsibility for this decision to the different regions of the image. Which regions have contributed most significantly to cancer prediction?
There are several stakeholders interested in answering this question, namely: the AI specialists who develop and train the models and the end-users of the model, for example, radiologists. The same interpretation of the decision is not always equally useful for all stakeholders. Therefore, it is necessary to adapt the techniques of interpretable AI to those who will take advantage of this information. For example, visual information may not be sufficient, and other forms of explanation are useful for a better understanding of the decision-making process, such as a descriptive text or a set of similar examples.
Currently, these interpretations, although still relatively basic, are already useful for diagnosing and improving the AI algorithm itself. If the interpretation highlights areas outside the lung in the image as relevant to a diagnosis of pneumonia based on an x-ray, probably the AI algorithm, even if it has decided well, will have "reasoned" poorly. If the analysis of an AI algorithm to support the recruitment of staff reveals that it is favouring men over women, there is a bias that needs to be removed.
The Workshop on "iMIMIC - Interpretability of Machine Intelligence in Medical Image Computing", which we organized last October 4th as part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) revealed the dynamics of the area, the potential and the virtues of these approaches. However, it also revealed the limitations and the much that remains to be done. The explanations/interpretations themselves are generated by an automatic AI algorithm, which has its own limitations and flaws. Explanations can also be manipulated. For example, an algorithm can use the customer's origin as a characteristic that conditions the granting of credit, but not use it to explain the decision. In another direction, it is essential to generalize explanations to cases beyond classification. How to explain that the estimated value for the sale value of the house is 437.52 and not just any other value? What is the proper explanation in this case? In yet another direction, how to explain a decision supported simultaneously by multiple sources of information (audio, text, video)?
The area of AI interpretability is taking its first steps. We still have a long way to go, to be followed with optimism, step by step. This joint progress, either focused on improving the decision, either on explaining the decision, is allowing for mutual growth of solutions for both tasks, in which we all win. It is not utopian to want to learn from the machine.