A more rigorous approach to xAI

David Tuckey, Alessandra Russo, Krysia Broda
Department of Computing, Imperial College London

June 21, 2019

The field of Explainable Artificial Intelligence (xAI) has boomed in the last couple of years with this term appearing in nearly all major AI conferences. This popularity has led to a huge number of publications and has made it harder and harder to summarize current research lines. There is nevertheless a focus on explaining so-called black-box models, such as neural networks, by generating a saliency map (highlighting the importance of the input features) or extracting knowledge from the model [3]. Overall we do not yet see the way forward that will allow explainable AI to have the impact many expect it will have on society. This might come from the fact that we lack clear and consensual definitions for the most fundamental terms that we use: *explanation, interpretation, explainability, interpretability*; meaning that we wrongly rely on them to justify our research. We focus on improving xAI techniques, but lack the tools to measure what an improvement is.

Bringing explanations to society is not an exclusivity of xAI; science has long been answering questions about physical phenomena. Taking inspiration from it, we argue that in order to aim towards a tool for measuring improvements in xAI techniques a stronger sense of rigour is needed. To do so the question might very well be "how can we deviate explainable AI off the engineering and technical path that it seems to be on ?" and consider it more as a science where extensive theoretical considerations are required.

With that in mind, we should start by asking ourselves the question: what is it we want to explain? What is the goal of our explanations? Without falling into the trap of using generic expressions like "improving trust" or "make interpretable", we can try defining explainable AI in terms of the impact it has on the user. Who we consider the explainee to be and what is the aim of the process are two of many important characteristics that we need to make explicit. An explanation intended for a child should be thought of differently from one intended for a machine learning engineer. Similarly, an explanation justifying a prediction can be different from one explaining the inference process. This would allow us to measure and classify approaches to xAI by their practical impact on people rather than by their technical characteristics. This may also help us relate our work to other disciplines which have been working on explanations for far longer than we have.

Integrating work from other disciplines might help xAI bring better explanations faster. We have at our disposal work in philosophy, physical sciences, social sciences, HCI, neuroscience and other fields where explanations play an important role. An example is highlighted by Miller in his review of explanations in social sciences and what it could bring to explainable AI [1]. Miller, among other things, presents research about the cognitive biases that humans have in their day-to-day interactions. We should consider these when creating xAI systems as we might be able to take advantage of them, or at least make sure they don't interfere with our explanations. This is just one example of what collaboration with other fields can bring.

One might think of a protocol to develop an explainable AI system that allows the inclusion of work from other disciplines and where the technical implementation is not the focus of the operation anymore, but rather is only one element in a broader scheme :

- Preliminary: What do we wish to explain? Who is the explainee? What impact should our explanation have?
- Study of the task: What kind of information do we want to extract from the system? How would a human explain it? What biases and methods would humans use to solve the task? How should we deliver the explanation to the explainee?
- Improvement of the technique: Build on top of current technical methods to extract the needed elements to build the explanation
- Verify the explanation: Verify with experiments that our explanations are relevant to the system's computation.
- Check the impact: Conduct a study to check that the explanations do have the expected impact on the target explainee.

The point of view of focusing explainable AI on the people rather than on the technique is starting to gain popularity and very good arguments are being made to back it up [2]. All in all, it can be seen as a very promising way forward. Much like we imitated birds to create planes, burdock burrs to create Velcro, we might want to take inspiration from humans to create explanations.

References

[1] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2017.

[2] Mireia Ribera and Agata Lapedriza. Can we do better explanations? A proposal of user-centered explainable AI. *CEUR Workshop Proceedings*, 2327, 2019.

[3] Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. (Nips), 2018.