Explainable Artificial Intelligence (XAI) in the humanities

This workshop is my first introduction to the term XAI and, as I understand it, it refers to AI that can be trusted and easily understood by humans [1] and is the opposite of the term 'black box', which refers to an object, or in this case an AI application, that has an input and an output, but there is no understanding of the internal workings of it [2]. I may have never used the term XAI during my work, but the experience I have had with applying AI has been very much in line with the ideas behind XAI. I am from a digital archaeology background and during one of the projects that I worked on we experimented with increasing the discoverability of zooarchaeological terms in unpublished archaeological reports using a Named Entity Recognition (NER) tool. One of the most important parts of the project was the relevancy of this tool to the end user, in our case the zooarchaeologist. This relevancy was achieved by explaining and making the tool understandable to zooarchaeologists during every step of the development of the tool. The term 'relevancy' is very much in line with XAI and a important part of both terms is transparency.

I have a number of concerns regarding XAI and I would like to explore these concerns from three different perspectives; the computer scientist or AI specialist developing these tools, the humanities scholar or humanities professional involved in the development of these tools or the one implementing these tools, and the end user at the completion of these projects. As I have a mixed background in both computer science, archiving and archaeology, I have been part of all three perspectives and have seen where potential problems may lie.

Starting with the computer scientist or developer of these tools. Machine learning is one of the most adapted techniques in AI. This machine learning, or deep learning, consists out of several layers to train an application to do some task, an example of this would be image classification. However, these layers are complicated and are barely understood by the developers themselves, some even referring to it as 'alchemy' [3]. Many computer scientists are more focused on the performance of the tool, than the actual understanding of what the tool is doing [4]. And this is becoming more and more of a problem, as we are building more complicated tools and distancing ourselves even further for the internal workings of them. There has been a change with the rise in popularity of XAI, one of my favourite examples being the AI detectives, who have developed a neural network to explain the internal working of other neural networks, which is quite ironic in some way [5]. But how should we explain the internal workings and be more transparent towards the end user when we are unsure ourselves what the tool is doing? And how do you ensure trust in these systems when you don't really know what is going on?

The next point regards the humanity scholar or professional who helps during the development of these tools or implements them in certain points of their workflow. Online, many articles can be found on the demystifying of AI and especially machine learning or deep learning [6]. This indicates that there is a consensus among many users that these tools are complicated and hard to implement. A lot of humanities scholar are not aware of how these tools operate and if they are not aware, how should these tools then be translated in a human explainable way to the end user? People within the humanities field should not become savvy data scientists, but they should become more aware of the benefits and drawbacks of these tools. It is not about being able to develop these tools, what is of importance is the critical analysis of these tools and understanding the impact they could have on research projects and institutions. If humanities professionals and scholars understand this, it will be possible for them to explain these tools in a better manner to a wider audience.

The last concern I have is around the transparency towards end users. In both the archiving and archaeological field that I have worked in, this is an important part of developing these tools. A lot of them are developed with end users in mind. But how should this transparency be mapped to users? A lot of writers are warning us about the dangers of seeing these tools as objective, such as Noble and O’Neill [7], [8]. But how do we map out the human decision making that occurred when creating these tools, especially when regarding above mentioned points? Also, how do we map out the probability, the chance that these tools are correct? Say an application is 60% sure a term should be tagged as ‘dog’ and 65% sure the term should be tagged as ‘cat’, how should this probability be mapped to a user in a human explainable and transparent way?

I have no answers to any of the above questions, but they are some real concerns that I have regarding XAI. In my opinion, and especially within the fields that I work in, before we start explaining and making AI transparent to the general public that we work with, we need to understand the systems ourselves to a certain extent and also critically analyse these tools on the relevancy they have for our practices. This should not only be the responsibility of the developer of the tools, but also the person implementing them into their systems.

References

[1] O. Biran and C. Cotton, ‘A Framework for Explanation of Machine Learning Decisions’, in *Workshop on Explainable AI (XAI)*, Melbourne, Australia, 2017, pp. 8–13.

[2] M. Bunge, ‘A General Black Box Theory’, *Philosophy of Science*, vol. 30, no. 4, pp. 346–358, 1963.

[3] A. Rahimi and B. Recht, ‘Reflections on Random Kitchen Sinks’, *Arg min blog*, 2017.

[4] K. L. Wagstaff, ‘Machine Learning that Matters’, presented at the Twenty-Ninth International Conference on Machine Learning (ICML), Edinburgh, 2012, pp. 529–536.

[5] P. Voosen, ‘The AI detectives’, *Science*, vol. 357, no. 6346, pp. 22–27, 2017.

[6] B. Dickson, ‘Tag: Demystifying AI’, *Techtalks*. [Online]. Available: https://bdtechtalks.com/tag/demystifying-ai/. [Accessed: 14-May-2019].

[7] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

[8] C. O’Neill, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin, 2016.