Andrea Berber, Nenad Filipovic, Sanja Sreckovic

Explanations in AI: The Anthropocentric 'Why?' versus Fidelity

Decision-making systems based on machine learning (ML) algorithms are being increasingly used to make decisions that significantly affect people's lives (e.g. medical diagnoses, loan qualification, etc.). Practical, ethical and legal concerns related to the opaqueness of ML algorithms impose the need to provide transparency by explaining the algorithms' working, known in the literature as the 'Explainable AI (xAI)' project.

The term 'explanation' is very broadly understood in xAI, and refers to any tool that provides some insight into the operation of the algorithm. This term is not explicitly defined, but it is clear that the explanations offered often diverge from both everyday and scientific notions of explanation, thus the issue of adequacy of the term 'explanation' in this context arises.

We argue that these problems arise because of a discrepancy between the notion of explanation implicit in the requests made from outside of the ML community (e.g., regulatory legislations), and the explanations offered in the xAI community. We claim that the former notion aims to answer what we term 'the anthropocentric Why?', and clashes with the detailed technical explanations of the inner workings of the ML model, which rarely seem explanatory to end-users. Alternatively, the post-hoc explanations of particular decisions may seem convincing to the end-users, but can often be simplistic and misleading. Thus there is a trade-off between explanation fidelity and the psychological persuasiveness for the end-user, and this trade-off leads to a methodological dilemma: to either accept potentially misleading explanations, or abandon anthropocentric notion of explanation.