ACAI 2021
The advanced course on AI on Human Centered AI

Tutorial Outline

Why and how should we explain in AI?
Stefan Buijsman, TU Delft

1. Why XAI?
As stated in the introduction, I think a nice starting point for the more conceptual discussion is to look at why XAI is relevant in the first place, now that there is a debate about not using black-box algorithms (Rudin, 2019) and, conversely, about using black-box algorithms without XAI (Robbins, 2019; Duran, 2021 has a more nuanced view but does argue that transparency is not necessary, as long as the AI is reliable). The central question here should be: what is the role of XAI? I will steer towards the idea that it is required for users to make informed decisions based on the AI output. Yet, for that kind of argument to work, we also need to say what information XAI should offer, for it to fulfill its role. For that, I turn to different definitions of 'explanation', though it's good to make sure all the participants are familiar with existing XAI tools (since I plan to relate the theoretical definitions to these current tools).

2. Technical tools/definitions in XAI
A very brief overview of approaches current in XAI, focusing on the central concepts more than the exact technical implementation, though I aim to highlight difficulties with each approach. For example: for the counterfactuals (Wachter et al, 2018)/algorithmic recourse (Karimi et al, 2021) tools in XAI challenges in finding an appropriate distance function, as well as selecting the most relevant counterfactual will be highlighted. Brief discussions of SHAP, LIME, saliency maps, so that students are aware of the different approaches to explaining AI taken in the literature, just to make sure everyone is on the right page for 2 and 3.

3. Conceptual approaches to explanation
XAI methods are, occasionally accompanied by a view on what 'explanation' is. Wachter et al (2018), for example, provide one. Similarly, Watson and Floridi (forthcoming, 'The explanation game') give a formal framework for what XAI should be providing. I plan to complement these discussions stemming from XAI with the more general discussion from philosophy (of science) on explanation, in the spirit of Miller (2019) though with a more thorough look at the different definitions extant in the literature. Durán (2021) does so from the unificationist account of explanation (cf. Kitcher (1989). Woodward (2003) offers a competing account, which acted as inspiration to Watson and Floridi, though they diverge from it somewhat. Mechanists (e.g. Kaplan, 2015) give a third account of explanation. All of these can be applied in the XAI context, giving different perspectives on what features characterize an explanation and so on what technical tools should focus on. For example, Woodward stresses counterfactuals, whereas the other two accounts do not actively include these in their characterizations of explanation.

4. HCI/Empirical evaluations of explanations

To avoid the tutorial from being too purely conceptual, abstracted away from what is currently (technically) feasible I plan to devote the remaining time to a different approach to characterizing (good) explanations, namely the empirical evaluation of different XAI methods. These provide more hands-on insight into the practical workings of XAI tools (e.g. the difficulty users have in generalizing from a list of SHAP explanations, Chromik et al, 2021 or saliency maps, Alqaraawi et al, 2020) and how these limitations might be anticipated from the conceptual work on explanation discussed in the first parts of the tutorial. Aside from concrete examples that illustrate these points, the aim is also to look at how XAI tools are evaluated and, in that way, at what the goal is of providing explanations (e.g. simulatability, Hase & Bansal, 2020; trust calibration, Wang & Yin, 2021). The final aim of the tutorial is to give students, through this combination of HCI and more conceptual work, a good overview of existing approaches to XAI and the characteristics of good explanations. Together that will provide a sense of what is still missing in XAI tools before we can fully explain AI systems, and why it is important to work on this topic.