Explaining AI to All – Towards More Comprehensible Communication via Explainability Labels

Abstract

The impact of artificial intelligence (AI) on the real world has increased significantly. The wide-spread use of highly complex models including language models like ChatGPT has changed how people react to and interact with AI systems. However, models of such complexity are de-facto black-boxes, meaning that their internal workings and decision making processes are incomprehensible even to experts in the field. As a result, a new research direction now brings forth designated methods to *explain* AI systems and their decisions.

Investigations of explainablity, in general, can be grouped into two different groups. Firstly, certain AI methods and resulting models are deemed inherently understandable, meaning that reasons for decisions are comprehensible per design. Secondly, for models that do not provide inherent explanations, external methods can be used to explain the model behavior in a post-hoc manner. Both inherent and external explanations can be further grouped into either describing global model behavior, or local phenomena for a specific (set of) decisions. However, with all mentioned explanation types in mind, the insight they provide is hard to understand for non-experts. As an example, consider feature attribution methods like Shapley values, which explain decisions by assessing the impact of each input feature via assigning positive and negative values. Without a strong understanding of method, these values provide little insight to non-experts and might even lead to misinterpretations about the decision process. Thus, we claim that in order to increase trust, we require a more comprehensible form of communicating about explainability aspects, that in addition to scientists also informs stakeholders unfamiliar with the subject matter, be it in academia, industry or the general public.

For reporting on other important properties of AI systems like energy efficiency and robustness, solutions were already proposed in literature. Drawing an analogy to the EU energy labels, recent work suggested a framework to convey information on resource trade-offs in a more comprehensible way. Their intricate reporting and labeling techniques allow to bridge the aforementioned communication gap in order to inform stakeholders of all knowledge levels. Despite the widespread call for more trustworthy and explainable models, we still lack a similar framework for understandable communication about these aspects. We propose to address this problem by pushing the labeling concept further, and adapt it for the domain of explainable AI. With our prototypical explainability labels, we demonstrate how both static properties as well as empiric behaviour can be communicated to

non-experts. Labels can be assembled to either describe inherent explainability aspects of ML methods, as well as reporting on external methods that allow for decision understanding. Our proposed framework can benefit developers who are not (yet) skilled in AI, and can now understand complex aspects of explainability at a more comprehensible level. Looking into the future, our system could be build upon by a certification authority, which awards labels that certify explainability and hence can improve trust into AI systems.

Keywords: explainability, reporting, certification, labeling