Critical Review: Causability and explainability of artificial intelligence in medicine

As research in the field of explainable AI grows, *Holzinger et al.* introduces the notion of *causability* to better understand the standard of explainability required in AI problems in the field of medicine. This paper is an intricate review of the current standard of explainable AI models (especially in deep learning), and the requirements to achieve a worthwhile level of explainable medicine. *Holzinger et al.* consider a useful explanation achieved from AI models to be similar to what professional colleagues in medicine exchange with each other. This is the focus of the paper, and the authors believe that conventional explainability notions have to be augmented by causability. Causability is defined as the measure of causal understanding achieved by a medical professional. This advocates the need for an explanation interface that *elaborates* the results of an explainable model to intuition or insights that can be utilized professionally. Moreover, the authors revisit general advancements in explainable AI, with an emphasis on deep learning. Also, differences between *ante-hoc* and *post-hoc* explainable models are discussed, while merits and issues of each of the type are explained. This is coupled with an example of both types on analysis done by a professional in liver pathology. The authors realize the need of concrete explainability (termed as causability) for optimal human-AI interaction, and thus propose the following measures: causability be established as a new scientific field, and weakly supervised learning algorithms be researched to cater for uncertainties involved with *ground truth* in the medical field.

This paper is an in-depth review on the state of the evolution of explainable AI, from the inception of AI (when models were inherently explainable owing to the rule-based approach) to the era of deep learning (when models containing millions of parameters offer little interpretability). The background information is extensive with detailed discussion on the types of uncertainty, and the types of explainable models, without losing the medical perspective. The paper reasonably realizes the pace of research in deep learning models, and their adverse effects on human-AI trust in the medicine industry. Given that, it argues the merits of current research and explainable AI to naturally build upon the notion of causability. To augment the extensive scientific review, the authors add a human perspective by introducing explanations by a medical professional. This coupled with un-biased rhetorical questions provide readers with an impartial view of the current situation of AI in medicine along with a way forward.

The paper proposes a formal definition of causability and relates it as a link to convert partially meaningful results from an explainable model to concrete findings via an interface. This is both feasible and practical, as it allows for mutual knowledge incorporation (professionals can gain insight from crunching diverse data, while they can also impart insight to the model i.e., if the explanation interface allows it). Currently explainable models reveal rather cryptic results, that are specific enough to explain predictions but far from what two medical professionals might share, fathom, or exchange. Dedicated research in causability will bridge the gap to allow for increased human-AI trust and ultimately better diagnoses. After analyzing the post-hoc and ante-hoc explanation provided by a trained medical profession, the paper proposes weak supervised learning algorithm to reduce the burden of establishing concrete *ground truths*. This establishes a much needed compromise that also has the advantage of being fault tolerant to uncertainties in true labels.

While the propositions mentioned in this paper aim to alleviate human-AI trust issues in the medical field, by making interpretations more understandable, it does not address the fact that these trust issues may have deeper stems. Data protection, privacy, limited resources, and socio-economic issues are a huge barrier to AI meaningfully assisting medical professionals. Furthermore, the suggested methods require a high level of cooperation by medical professionals for accurate annotations; and even with accurate annotations, evaluating the success of explainable AI models is also manual. This is worse in the case of evaluating causability, as causability is something that is defined from the perspective of comprehension by a person.

This paper reviews that generally prediction performance decrease with increase in explainability of a model. This aspect of explainability may have adverse effects on the goal of maximizing human-AI trust, which has not been adequately discussed in this paper. In worst case, it may also result in contradicting insights that may negatively affect a diagnoses. The aftermath of such events may propagate considering the sensitivity of the medical field.

To conclude, while there are concerns regarding the inclusion of AI in medicine, the benefits of a properly implemented human-AI system far outweigh the risks. However, the authors realize the extend of such a task, and propose middle grounds for sustainable growth. Their in-depth literature review along with a scientific approach to defining causability show promising directions for future research.