From Explainable AI to Human-Centered AI

Andreas Holzinger

Medical University Graz, Austria
andreas.holzinger@human-centered.ai

Abstract. The problem of explainability is as old as AI itself and classic AI represented comprehensible retraceable approaches. Their weakness was in dealing with non-linearities and the intrinsic uncertainties of medical data. Advances in data-driven statistical machine learning have led to the current renaissance of AI, but the solutions are becoming increasingly complex and opaque. Due to increasing social, ethical, and legal aspects of AI in medicine, explainable AI (xAI) is attracting much interest within the international research community. While xAI deals with the implementation of transparency and traceability of statistical black-box machine learning methods, there is a pressing need to go beyond xAI, e.g. to extent explainability with causability. The integrative backbone for this approach is in interactive machine learning with the human-in-the-loop because a human domain expert complements AI with implicit knowledge. Humans are robust, can generalize from few examples, understand relevant representations and concepts and are able to explain causal links between them. Consequently, more research is needed on how human experts explain their decisions by examining their strategies, as they are (but not always) able to describe the underlying explanatory factors. Formalized, these can be used to build structural causal models of human decision making and characteristics can be mapped back to train AI. Finally, such an AI-ecosystem needs advanced Human-AI interfaces, that allow to ask questions of why, but also to ask for counterfactuals, i.e. what-if. This interactivity between human and AI will contribute to enhance robustness, reliability, accountability, fairness and trust in AI and foster ethical responsible machine learning with the human-in-control.