What is XAI anyway? Compared Findings and Thoughts on the Integration of Explainable AI in Interdisciplinary Empirical Research

2023-06-23 18:29:44

Authors

Jessica Szczuka, Nicole Krämer

Abstract

You would need to be a polymath to comprehend the growing complexities of everyday life. This includes understanding AI systems. Although humans encounter them every day (e.g., in terms of language models for various communicative purposes, recommender systems, or recognition systems), most people lack an understanding of the technology behind the black boxes. Consequently, users must rely on their trust in these systems (and their outputs) to still use them without constantly questioning it‘s usage. Nonetheless, researcher agree that some understanding of what is happening inside the black boxes is an important key to informed and empowered use of such systems. One popular approach is explainable AI (XAI), i.e., self-explanatory systems. What can be achieved by such explanations, what are the limitations, and how can the impact of explanations be adequately studied empirically? In this talk, three studies (Horstmann et al., 2023; Herter et al., 2023; Szymczyk et al., 2023) of our department will be presented in which the effect of XAI has been experimentally investigated. First overall conclusions can be drawn: (1) The effect of XAI on trust is context and relevance-dependent: While we found enhanced trust in the system based on explanations in the health context (here psychotherapy), this was not the case in less relevant areas (breakfast recommendation). (2) Knowledge transfer based on XAI depends on the explanation: the wording and therefore computational approach behind the explanation is crucial (e.g., nearest neighbor vs. counterfactual). (3) The potential of more ?social? explanations, meaning the usage of an artificial persona as a source of epistemic trust is yet to be investigated. (4) The design of empirical studies heavily determines the implications. To address the latter in detail, successes but also pitfalls for the interdisciplinary empirical study of XAI are highlighted. A. C. Horstmann, J. Szczuka, L. Mavrina, A. Artelt, C. Strathmann, N. Szymczyk, L. Michelle Bohnenkamp, and N. Krämer (2023). Enhancing the Understanding of Algorithms With Contrastive Explanations: An Experimental Study on

the Effects of Explanations and Person-Likeness on Trust in and Understanding of Algorithms, presented at annual meeting of International Communication Association (ICA)

N. Szymczyk, G. Gül, L. Klein, G. Schneider, F. Wenda, J. Zumbrägel M. Wischnewski and J. Szczuka (accepted). Trust Me, I'm AI: Examining the Influence of XAI and/or AI-Seal on User Trust and Understanding presentation at 3th Conference of the Media Psychology Division (DGPs)

E. Herter, T. Kühn, L. Mühl, A Schilly, L. Stecker, M. Wischnewski and J. Szczuka (accepted). XAI in initial psychotherapy consultation: An experimental study on the effects of XAI on clients' trust in their (virtual) therapist presentation at 3th Conference of the Media Psychology Division (DGPs)

Keywords

XAI, Trust, Understanding, Empirical Research