Explanations that backfire: Explainable artificial intelligence can cause information overload

Aidah Ferguson
University College London, London, United Kingdom

Matija Franklin
UCL, London, United Kingdom

David Lagnado
University College London, London, United Kingdom

Abstract

Explainable Artificial Intelligence (XAI) provides human understandable explanations into how AI systems make decisions in order to increase transparency. We explore how transparency levels in XAI influence perceptions of fairness, trust and understanding, as well as attitudes towards AI use. The transparency levels – no explanation, opaque, simple and detailed - were varied in two contexts - treatment prioritization and recidivism forecasting. In eight experimental groups, 573 participants judged these explanations. As predicted opaque explanations decreased trust and understanding, but surprisingly simple explanations that were more limited in the information they provided had stronger effects on trust and understanding than detailed explanations. Transparency levels did not have an impact on perceptions of fairness and attitudes towards AI, but context did, with the recidivism AI being perceived as less fair. The findings are discussed in relation to information overload and task subjectivity vs objectivity.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*