Good Explanation for Algorithmic Transparency

Joy Lu, Dokyun (DK) Lee, Tae Wan Kim, David Danks
tonglu@andrew.cmu.edu, dokyun@cmu.edu,
twkim@andrew.cmu.edu, ddanks@cmu.edu
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

ABSTRACT

Machine learning algorithms have gained widespread usage across a variety of domains, both in providing predictions to expert users and recommending decisions to everyday users. However, these AI systems are often black boxes, and endusers are rarely provided with an explanation. The critical need for explanation by AI systems has led to calls for algorithmic transparency, including the "right to explanation" in the EU General Data Protection Regulation (GDPR). These initiatives presuppose that we know what constitutes a meaningful or good explanation, but there has actually been surprisingly little research on this question in the context of AI systems. In this paper, we (1) develop a generalizable framework grounded in philosophy, psychology, and interpretable machine learning to investigate and define characteristics of good explanation, and (2) conduct a large-scale lab experiment to measure the impact of different factors on people's perceptions of understanding, usage intention, and trust of AI systems. The framework and study together provide a concrete guide for managers on how to present algorithmic prediction rationales to end-users to foster trust and adoption, and elements of explanation and transparency to be considered by AI researchers and engineers in designing, developing, and deploying transparent or explainable algorithms.

Link to full paper: https://drive.google.com/open?id=1-QkGGnW1EmprtvADGGmayQPv4E-L7U2b

CCS Concepts/ACM Classifiers

• Computing methodologies~Artificial intelligence
• Computing methodologies~Philosophical/theoretical foundations of artificial intelligence

Author Keywords

explainable AI, interpretable AI, lab experiments

AIES '20, February 7–8, 2020, New York, NY, USA.

ACM ISBN 978-1-4503-7110-0/20/02.
DOI: https://doi.org/10.1145/3375627.3375821