Explainable AI, Disagreement, and Idealization

Dr. Will Fleisher
Department of Philosophy
Center for Digital Ethics
Georgetown University

AI systems are being used for a rapidly increasing number of important decisions. Many of these systems are "black boxes": their functioning is opaque both to the people affected by them and to those developing them. Black box AI systems are difficult to evaluate for accuracy, fairness, and general trustworthiness. Explainable AI (XAI) methods aim to alleviate the opacity of complex AI systems. However, there is debate about whether these methods can provide adequate explanations for the behavior of black box AI systems. One of the biggest problems for XAI methods is that they are prone to disagree with one another. In this paper, I argue that we should understand these XAI methods as producing idealized models of black box systems, much like idealized scientific models of other complex phenomena. This account of XAI methods employs resources from epistemology and philosophy of science to help understand what it would mean for XAI methods to successfully explain black box AI systems. These resources can also help to partially alleviate the disagreement problem for XAI.

2:00pm
Fri 21 April

Dewey 2110-D

Sponsored by the Central New York Humanities Corridor from an award by the Mellon Foundation

Central New York
Humanities Corridor