Explanation, Justification, and Responsibility

In this talk, I will show how methods from explainable AI (XAI) can be used for the purposes of justification. To this end, I will introduce a distinction from philosophy of action between explanatory, motivating, and normative reasons, and show how (and to what extent) XAI methods can provide insight to these reasons within the context of AI-driven actions. Building on this account of justification, I will tackle the problem of responsibility gaps in autonomous and opaque AI. Here, I will argue that insofar as XAI methods allow us to predict, control, and justify an AI system's actions, they can also help us close any ensuing responsibility gaps.