Fair and Bias-Aware Algorithmic Assistance

Jann Spiess

Stanford Graduate School of Business

Abstract

When machine-learning algorithms are deployed in high-stakes decisions, they often assist human decision-makers rather than replace them. In this talk, I discuss how properties of machine predictions can affect the fairness and efficiency of resulting decisions when a biased human decision-maker retains the ultimate decision authority. In the first and main part (based on work with Talia Gillis and Bryce McLaughlin), I consider the effect of algorithmic design on the fairness of machine-assisted human decisions. I argue in a formal model that the inclusion of a biased human decision-maker can revert common relationships between fairness and accuracy. Specifically, I demonstrate that excluding information about protected groups from the prediction may fail to reduce, and may even increase, ultimate disparities. In the second part (based on work with Bryce McLaughlin), I briefly present a model of human–machine interaction in which the decision-maker does not only react to the information contained in an algorithmic recommendation. Instead, the human agent sees the recommendation as a default action that may be costly to deviate from. I demonstrate that such recommendation-dependent preferences create inefficiencies where the decision-maker is overly responsive to the recommendation, which changes the optimal design of the algorithm towards providing less conservative recommendations. As a potential remedy, I discuss an algorithm that strategically withholds recommendations, and show how it can improve the quality of final decisions. Overall, the results I present suggest that any study of critical properties of complex decision systems, such as the fairness and efficiency of machine-assisted human decisions, should go beyond focusing on the underlying algorithmic predictions in isolation, and that behavioral aspects of human decision-making should influence the design of algorithmic assistance.