The Measure and Mismeasure of Fairness

Hans Gaebler
Harvard University

The field of fair machine learning aims to ensure that decisions guided by algorithms are equitable. Over the last decade, several formal, mathematical definitions of fairness have gained prominence. Here we first assemble and categorize these definitions into two broad families: (1) those that constrain the effects of decisions on disparities; and (2) those that constrain the effects of legally protected characteristics, like race and gender, on decisions. We then show, analytically and empirically, that both families of definitions typically result in strongly Pareto dominated decision policies. For example, in the case of college admissions, adhering to popular formal conceptions of fairness would simultaneously result in lower student-body diversity and a less academically prepared class, relative to what one could achieve by explicitly tailoring admissions policies to achieve desired outcomes. In this sense, requiring that these fairness definitions hold can, perversely, harm the very groups they were designed to protect. In contrast to axiomatic notions of fairness, we argue that the equitable design of algorithms requires grappling with their context-specific consequences, akin to the equitable design of policy. We conclude by listing several open challenges in fair machine learning and offering strategies to ensure algorithms are better aligned with policy goals.