Katherine Strandburg

New York University

Decision-making, Machine Learning and the Value of Explanation

Much of the policy and legal debate about algorithmic decision-making has focused on issues of accuracy and bias. Equally important, however, is the question of whether algorithmic decisions are understandable by human observers: whether the relationship between algorithmic inputs and outputs can be explained. Explanation has long been deemed a crucial aspect of accountability, particularly in legal contexts. By requiring that powerful actors explain the bases of their decisions — the logic goes — we reduce the risks of error, abuse, and arbitrariness, thus producing more socially desirable decisions. Decision-making processes employing machine learning algorithms complicate this equation. Such approaches promise to refine and improve the accuracy and efficiency of decision-making processes, but the logic and rationale behind each decision often remains opaque to human understanding. Indeed, at a technical level, it is not clear that all algorithms *can* be made explainable and, at a normative level, it is an open question when and if the costs of making algorithms explainable outweigh the benefits. This presentation will begin to map out some of the issues that must be addressed in determining in what contexts, and under what constraints, machine learning approaches to governmental decision-making are appropriate.