Translation tutorial:
21 fairness definitions and their politics

Arvind Narayanan
(Computer scientist, Princeton University)

Computer scientists and statisticians have devised numerous mathematical criteria to define what it means for a classifier or a model to be fair. The proliferation of these definitions represents an attempt to make technical sense of the complex, shifting social understanding of fairness. Thus, these definitions are laden with values and politics, and seemingly technical discussions about mathematical definitions in fact implicate weighty normative questions. A core component of these technical discussions has been the discovery of trade-offs between different (mathematical) notions of fairness; these trade-offs deserve attention beyond the technical community.

This tutorial has two goals. The first is to explain the technical definitions. In doing so, I will aim to make explicit the values embedded in each of them. This will help policymakers and others better understand what is truly at stake in debates about fairness criteria (such as individual fairness versus group fairness, or statistical parity versus error-rate equality). It will also help computer scientists recognize that the proliferation of definitions is to be celebrated, not shunned, and that the search for one true definition is not a fruitful direction, as technical considerations cannot adjudicate moral debates.

My second goal is to highlight technical observations and discoveries that deserve broader consideration. Many of these can be seen as **"trolley problems" for algorithmic fairness**, and beg to be connected to philosophical theories of justice. I hope to make it easier for ethics scholars, philosophers, and domain experts to approach this territory.