

Algorithmic fairness

Group fairness

Goal (Dwork et al. 2012). *Similar individuals out to be treated similarly and operationalized it through an algorithmic framework.*

Group fairness

However, the notion of fairness is fuzzy. It has no precise meaning, nor does society agree on what it entails.

Question (Kleinberg-Mullainathan-Raghavan 2017). *What should we do when our ideas of fairness conflict?*

Original context

- *ProPublica* claimed to have uncovered racial bias in COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)'s algorithm, used in criminal justice.
- Developed by *Northpointe*, it helps judges decide whether to release defendants or incarcerate them.

Original context

- Based on a survey given to defendants, COMPAS assigns to each defendant a score between 1 and 10.
- Higher the score, more likely to be engaged in another criminal activity.

Original context

ProPublica's findings. Of those not rearrested, COMPAS has given Black defendants much higher scores than white counterparts.

Original context

Northpointe's response. COMPAS exhibited "accuracy equity" and "predictive parity": Whether rearrested or not, Black defendants received higher scores than white defendants. COMPAS exhibited similar predictive power for both.

Mathematical framework

- Consider the positive and negative classes of individuals (e.g., recidivists or non-recidivists).
- Design an algorithm for a risk assessment tool to predict the likelihood that an individual belongs to the positive class.

Mathematical framework

- Each individual comes from one of the two groups (e.g., Black and white defendants).
- Each individual is associated with a feature vector σ (e.g., a results of a survey).

Fairness conditions

S.1. Balance for the negative class.

Conditioned on belonging to the negative class, all have similar scores.

S.2. Balance for the positive class.

Conditioned on belonging to the positive class, all have similar scores.

ProPublica has found that COMPAS did not respect S.2 (nor S.1).

Fairness conditions

S.3. Calibration within groups. *For each group, and each predicted probability p , a p fraction of them belong to the positive class.*

Northpointe asserts that COMPAS satisfies S.3.

Impossibility theorem

Theorem 2.1 (Kleinberg-Mullainathan-Raghavan 2017). *Unless one of the following holds, S1, S2 and S3 cannot be simultaneously true:*

- *Everyone can be perfectly classified into the positive or negative class.*
- *The same fraction belongs to the positive class for both groups.*