The Measure and Mismeasure of Fairness

Presenters: Haoxian Chen and Chenkai Yu

Overview

- Two-part taxonomy of formal fairness definitions
- Statistical limitations of both families of fairness definitions
- Steps to build more equitable algorithms

Setting

- Obeserved covariates $X \sim \mathcal{D}_{\mathcal{X}}$ i.i.d.
- Discrete protected attributes $A = \alpha(X) \in \mathcal{A}$
- Binary decision $D \in \{0, 1\}$ determined by a rule $d(x) = \mathbb{P}\left(D = 1 | X = x\right)$
- Budget $\mathbb{E}[D] \leq b$
- Binary outcome Y and possibly two potential outcomes $Y(0)$ and $Y(1)$ affected by D

Examples

Diabetes screening

- Covariates (X): patient's age, body mass index (BMI), (race) etc.
- protected attribute (A): race
- Outcomes (Y): whether a patient has diabetes or not.
- Goal: design an equitable screening policy d to determine which patients to be screened, based on X.

Examples

College admissions

- Covariates (X): student's test score, (race) etc.
- protected attribute (A): race
- Causal outcomes (Y): $Y(1)/Y(0)$ describes whether an applicant would attain a degree if admitted/not admitted.
- Goal: design an equitable admission policy d to determine which students to admit.

Two-part Taxonomy of Fairness Definitions

Limiting the Effect of Decisions on Disparities

- Requires the policy to have equal error rates across groups, defined by protected attributes.

Limiting the Effect of Attributes on Decisions

- Limits the effect of protected attributes on policy decision.

Limiting the Effect of Decisions on Disparities

Demographic parity

$$D \perp\!\!\!\perp A.$$

Example(s):

- The proportion of patients who are screened for the disease is equal across race groups.
- An equal proportion of students is admitted across race groups.

Limiting the Effect of Decisions on Disparities

Equalized false positive rates

$$D \perp\!\!\!\perp A \mid Y = 0.$$

Example(s):

- The screening rates of individuals who in reality do not have diabetes are equal across race groups.

Limiting the Effect of Decisions on Disparities

Counterfactual predictive parity

$$Y(1) \perp\!\!\!\perp A \mid D = 0.$$

Example(s):

- In college admissions example, among rejected applicants, the proportion who would have attained a college degree, had they been accepted, is equal across race groups.

Limiting the Effect of Decisions on Disparities

Counterfactual equalized odds

$$D \perp\!\!\!\perp A \mid Y(1).$$

Example(s):

- among applicants who would graduate if admitted (i.e., $Y(1) = 1$), students are admitted at the same rate across race groups.
- among applicants who would not graduate if admitted (i.e., $Y(1) = 0$), students are again admitted at the same rate across race groups.

Limiting the Effect of Decisions on Disparities

Conditional principal fairness

$$D \perp\!\!\!\perp A \mid Y(0), Y(1), W.$$

Example(s):

- conditional principal fairness means that "similar" applicants are admitted at the same rate across race groups.

Limiting the Effect of Attributes on Decisions

Blinding

Suppose $\mathcal{X} = \mathcal{X}_u \times \mathcal{A}$, where $\mathcal{X}_u$ denotes "unprotected" attributes. Then blinding holds when for all $a, a' \in \mathcal{A}$ and $x_u \in \mathcal{X}_u$,

$$d(x_u, a) = d(x_u, a').$$

Example(s):

- The screening decision depends solely on factors like age and BMI.
- College admissions decisions depend only on factors like test scores and extracurricular activities.

Limiting the Effect of Attributes (Causal)

Counterfactual fairness

$$\mathbb{E}[D(a')|X] = \mathbb{E}[D|X],$$

where $D(a')$ denotes the decision when one's protected attributes are counterfactually altered.

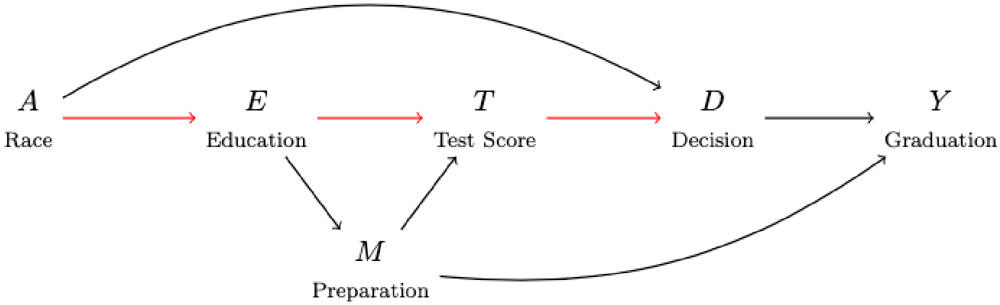
Example(s):

- For each group of observationally identical applicants (i.e., same values of X), the actual admitted proportion is the same as the proportion who would be admitted if their race were counterfactually altered.

Limiting the Effect of Attributes (Causal)

Path-specific counterfactual

Allows protected traits to influence decisions along certain causal paths but not others.



Limiting the Effect of Attributes (Causal)

Π-fairness

Let Π be a collection of paths, and, for a measurable function ω on $\mathcal{X}$, let $W = \omega(X)$ describe a reduced set of the covariates X. Path-specific fairness, also called Π-fairness, holds when, for any $a \in \mathcal{A}$,

$$\mathbb{E}[D_{\Pi,A,a'}|W] = \mathbb{E}[D|W],$$

where $D(a')$ denotes the decision when one's protected attributes are counterfactually altered.

Equitable Decisions without Trade-offs

Assume $\mathbb{E}[D] \leq b = 1$, i.e. no budget constraint.

Expected utility

Denote $v(y)$ by the benefit of making decision $D = y$ over $D = 1 - y$. Then the expected utility of making decision $D = 1$ over $D = 0$ is

$$u(x) := \mathbb{E}[v(Y)|X = x] = r(x) \cdot v(1) + [1 - r(x)] \cdot v(0),$$

where $r(x) := \mathbb{P}(Y = 1|X = x)$

Equitable Decisions without Trade-offs

Define the utility of a policy $\tilde{u}(d) := \mathbb{E}[d(X) \cdot u(X)]$.

Threshold policy maximizing utility

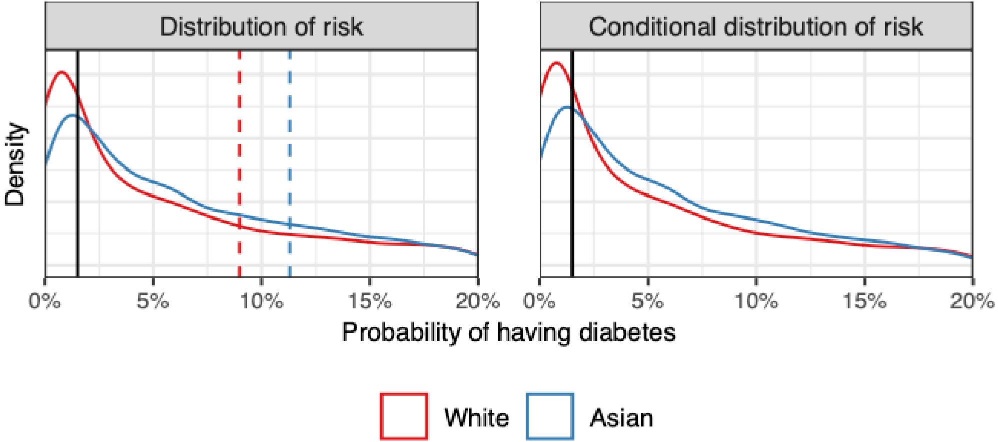
A decision policy $d^*(x)$ is utility-maximizing if $u(d^*) = \max_d \tilde{u}(d)$.
Therefore, we have

$$d(x) = \begin{cases} 1 & \text{if } r(x) > \frac{v(0)}{v(0)-v(1)} \\ 0 & \text{otherwise} \end{cases}$$

We call $t := \frac{v(0)}{v(0)-v(1)}$ as the optimal threshold.

The Problem of Classification Parity

However, threshold policies in general violate classification parity.

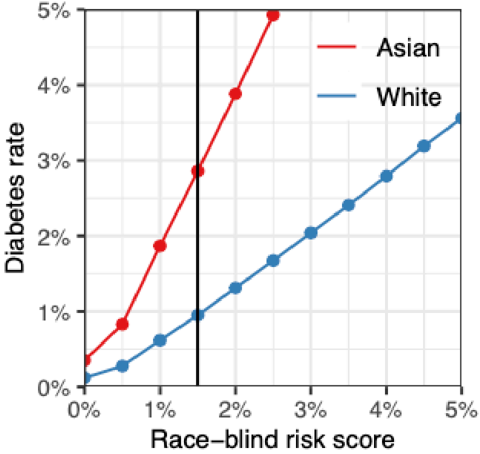


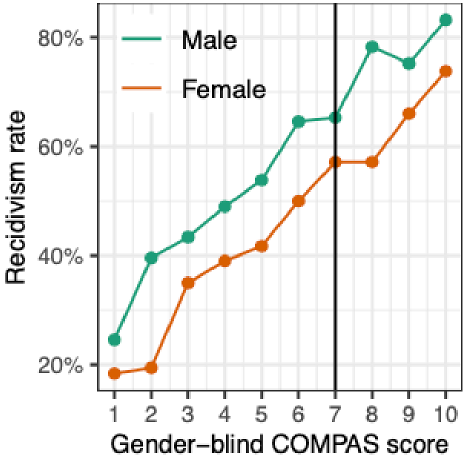
The Problem of Inframarginality

Theorem 9

If $0 < t < 1$, then for almost every collection of group-specific risk distributions which have densities on $[0, 1]$, no utility-maximizing decision-policy satisfies demographic parity or equalized false positive rates.

The Problem with Fairness through Unawareness





The Problem with Fairness through Unawareness

Theorem 10

Suppose $0 < t < 1$, where t is the optimal decision threshold on the risk scale, as in Eq. (9). Let $\pi : \mathcal{X}_u \times \mathcal{A} \to \mathcal{X}_u$ denote restriction to the unprotected covariates. Let $\rho(x) = \Pr(Y = 1 \mid \pi(X) = \pi(x))$ denote the risk estimated using the blinded covariates. Suppose that $r(x)$ and $\rho(x)$ have densities on $[0, 1]$ that are positive in a neighborhood of t. Further suppose that there exists $\epsilon > 0$ such that the conditional variance $\mathrm{VAR}(r(X) \mid \rho(X)) > \epsilon$ a.s., where $r(x)$ is the risk estimated from the full set of covariates. Then no blind policy is utility-maximizing.

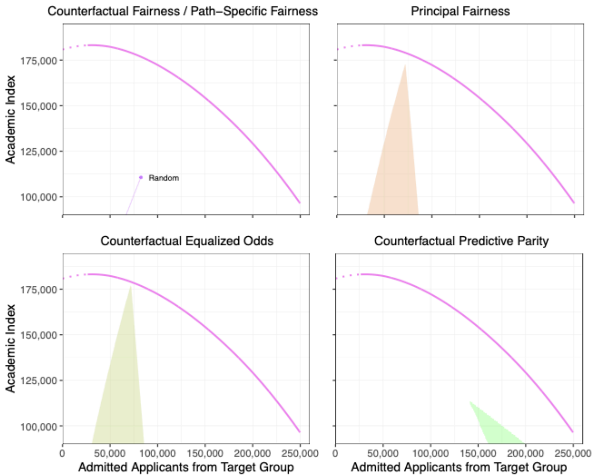
Equitable Decisions in the Presence of Trade-offs

Multi-objective

In the setting of $b < 1$, we need consider the tradeoff between two competing objectives:

$$u_1(d) := \mathbb{E}[m(X) \cdot d(X)], \quad u_2(d) := \mathbb{E}[\mathbb{1}_{\alpha(X)=\alpha_1} \cdot d(X)]$$

The Geometry of Fair Decision Making



Theory of Fairness in the Presence of Trade-offs

Consistency of utility

We say that a set of utilities $\mathcal{U}$ is consistent modulo α if, for any $u, u' \in \mathcal{U}$:

1. For any $x, \operatorname{sign}(u(x)) = \operatorname{sign}\left(u'(x)\right)$;
2. For any x_1 and x_2 such that $\alpha\left(x_1\right) = \alpha\left(x_2\right), u\left(x_1\right) > u\left(x_2\right)$ if and only if $u'\left(x_1\right) > u'\left(x_2\right)$.

Theory of Fairness in the Presence of Trade-offs

Pareto dominance

Suppose $\mathcal{U}$ is a collection of utility functions.

- Pareto dominated: For a decision policy d, there exists a feasible alternative d' such that $u\,(d') \geq u(d)$ for all $u \in \mathcal{U}$, and there exists $u' \in \mathcal{U}$ such that $u'\,(d') > u'(d)$.
- Strongly Pareto dominated: For a decision policy d, there exists a feasible alternative d' such that $u\,(d') > u(d)$ for all $u \in \mathcal{U}$.
- Pareto efficient: A policy d that is feasible and not Pareto dominated.
- Pareto frontier: The set of Pareto efficient policies.

Theory of Fairness in the Presence of Trade-offs

Characterization of Pareto efficient policy

Suppose $\mathcal{U}$ is a set of utilities that is consistent modulo α. Then any Pareto efficient decision policy d is a multiple-threshold policy. That is, for any $u \in \mathcal{U}$, there exist group-specific constants $t_a \geq 0$ such that, a.s.:

$$d(x) = \begin{cases} 1 & u(x) > t_{\alpha(x)} \\ 0 & u(x) < t_{\alpha(x)} \end{cases}$$

Theory of Fairness in the Presence of Trade-offs

$\mathcal{U}$-fineness distribution

Let $\mathcal{U}$ be a collection of functions from $\mathcal{Z}$ to $\mathbb{R}^d$ for some set $\mathcal{Z}$. We say that a distribution of Z on $\mathcal{Z}$ is $\mathcal{U}$-fine if $g(Z)$ has a density for all $u \in \mathcal{U}$.

Limitations of Fairness Definitions

Theorem 17

Suppose $\mathcal{U}$ is a set of utilities consistent modulo α. Further suppose that for all $a \in \mathcal{A}$ there exist a $\mathcal{U}$-fine distribution of X and a utility $u \in \mathcal{U}$ such that $\Pr(u(X) > 0, A = a) > 0$, where $A = \alpha(X)$.

Then

- For almost every $\mathcal{U}$-fine distribution of X and $Y(1)$, any decision policy satisfying counterfactual equalized odds is strongly Pareto dominated.
- If $|\mathrm{Img}(\omega)| < \infty$ and there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(A = a, W = w) > 0$ for all $a \in \mathcal{A}$ and $w \in \mathrm{ImG}(\omega)$, where $W = \omega(X)$, then, for almost every $\mathcal{U}$ fine joint distribution of $X, Y(0)$, and $Y(1)$, any decision policy satisfying conditional principal fairness is strongly Pareto dominated.
- If $|\mathrm{ImG}(\omega)| < \infty$ and there exists a $\mathcal{U}$-fine distribution of X such that $\Pr(A = a, W = w_i) > 0$ for all $a \in \mathcal{A}$ and some distinct $w_0, w_1 \in \mathrm{IMG}(\omega)$, then, for almost every $\mathcal{U}^{\mathcal{A}}$-fine joint distributions of A and the counterfactuals $X_{\Pi,A,a'}$, any decision policy satisfying path-specific fairness is strongly Pareto dominated. [14]

Limitations of Fairness Definitions

Corollary 18

Consider a utility of the form

$$u^*(d) = v\left(\mathbb{E}[m(X) \cdot d(X)],\ \mathbb{E}\left[\mathbb{1}_{\alpha(X)=a_1} \cdot d(X)\right]\right)$$

where v is monotonically increasing in both coordinates and $m(x) \geq 0$. Then, under the same hypotheses as in Theorem 17, for almost every joint distribution, no utility-maximizing decision-policy satisfies counterfactual equalized odds, conditional principal fairness, or pathspecific fairness.

Ways to Improve Equitability

Balancing inherent trade-offs in decision problems

- Explicitly calculate the Pareto frontier.
- If not possible to compute, list and discuss trade-offs to reduce the risk of adopting problematic policies, like those satisfying some formal fairness criteria.

Ways to Improve Equitability

Assessing calibration

- Check whether risk scores correspond to the same observed level of risk across groups.
- Measure calibration: regress observed outcomes against risk estimates and group membership.
- Rectifying miscalibration:
    - Training group-specific models
    - Include group membership in a single model
    - Include additional non-protected covariates

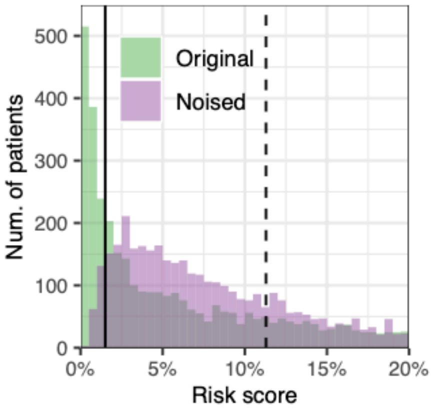
Calibration is not sufficient

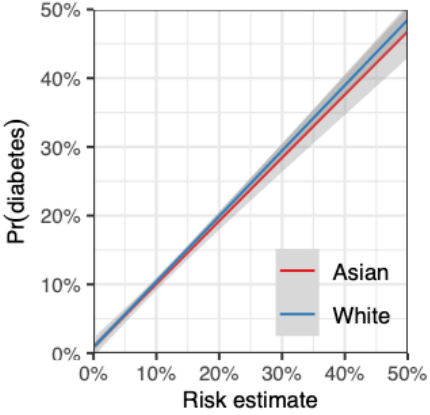
Loan application

- Within zip code, White and Black applicants have similar default rates.
- Black applicants live in zip codes with relatively high default rates.
- The bank would tend to refuse Black applicants.

Calibration is not sufficient

Diabetes example with noisy covariates





Ways to Improve Equitability

Selecting the targets of prediction

- Label bias: a mismatch between our true outcome of interest and the available data.
  - Heavier policing in communities of color might lead to Black and Hispanic defendants being arrested. Data might cause underestimation of the risk posed by White defendants.
  - Suppose counterfactual outcome is of our interest, we cannot observe it in reality (e.g., release vs detention).
- One way to mitigate label bias is to adjust the target of interest.
  - To represent one's medical need, health status could be a better proxy than medical cost.

Ways to Improve Equitability

Collecting training data

- Ideal: datasets are representative of the populations on which they are ultimately applied.
- Value: it depends on the degree to which race, gender and other protected attributes are predictive.
- Benefits:
  - At training, full support of features is present.
  - At model validation, representative samples helps to assess the model's generalization.