The Alan Turing Institute · NYU

Counterfactual Fairness

Matt J. Kusner*[1,2], Joshua R. Loftus*[1,3], Chris Russell*[1,4], Ricardo Silva[1,5]
*Equal Contribution [1]The Alan Turing Institute [2]University of Warwick [3]New York University [4]University of Surrey [5]University College London

ML is making Life-Changing Decisions

Trump's Proposed Extreme Vetting Software [Biddle, 2017]


NATIONAL SECURITY INVESTIGATIONS DIVISION (NSID)
Vetting of the Future:
• Comprehensive utilization of innovative big-data programs to enhance current vetting procedures and processes.

We Have Big Problems

Our Solution

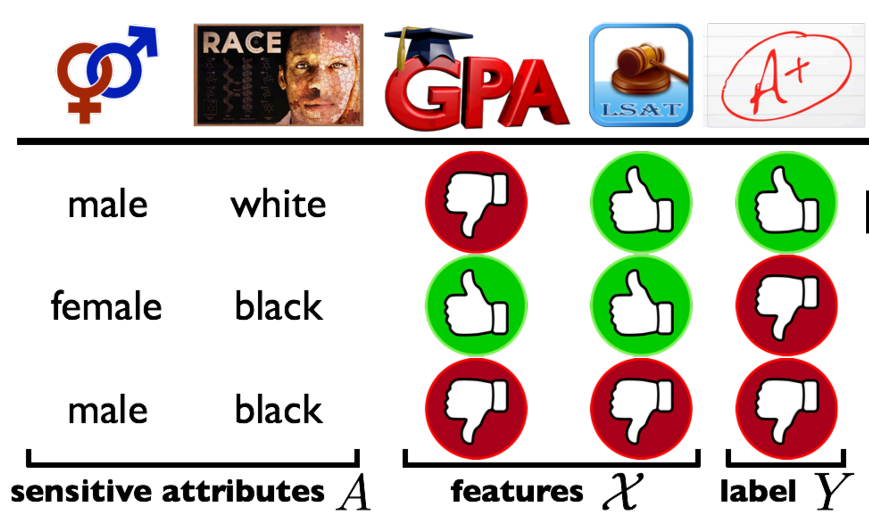
A **fair classifier** gives the **same prediction** had the person **had a different race/sex**.

We present a metric to check if **any algorithm is fair**

And an algorithm to **learn fair classifiers**

Law School Admissions

Data **Goal**



$\hat{Y} : \mathcal{X}, A \longrightarrow Y$

Prior Fairness Methods

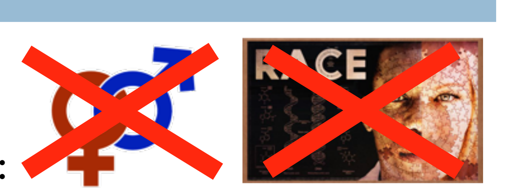
Fairness through Unawareness [Grgić-Hlača et al., 2016]

Equality of Opportunity [Hardt et al., 2016]

Prior Fairness Definitions Are Insufficient

Fairness through Unawareness [Grgić-Hlača et al., 2016]

remove sensitive attributes:



$\hat{Y} : \mathcal{X} \longrightarrow Y$

features $\mathcal{X}$ are biased

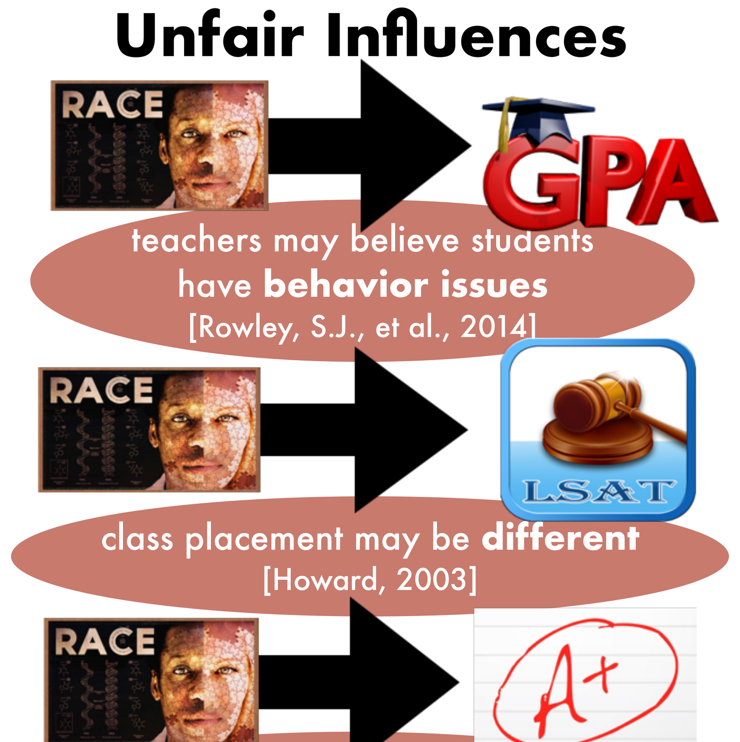
Equality of Opportunity [Hardt et al., 2016]

black

$\mathbb{P}(\hat{Y}=1 \mid A=a, Y=1) =$
$\mathbb{P}(\hat{Y}=1 \mid A=a', Y=1)$

white

Unfair Influences



teachers may believe students have **behavior issues** [Rowley, S.J., et al., 2014]

class placement may be **different** [Howard, 2003]

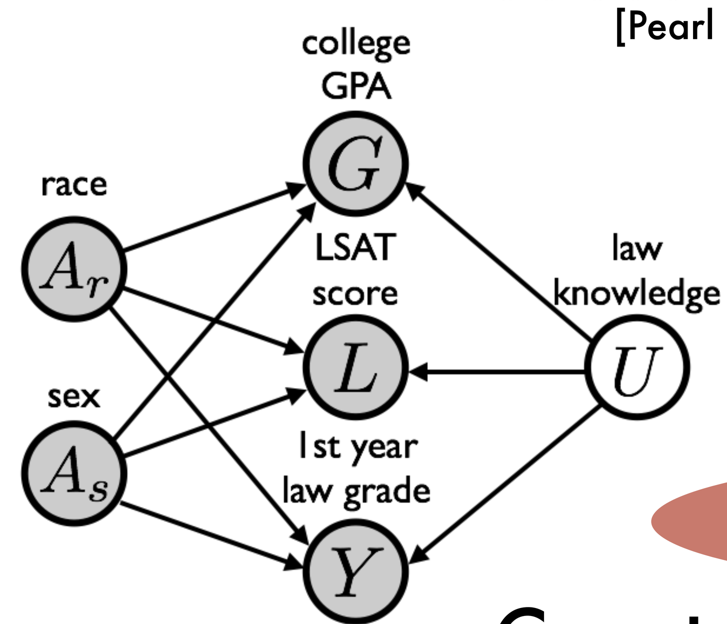
lack of minority race teachers **affects minority student outcomes** [Birdsall et al., 2016]

35% of black students
50% of white students

label Y is biased

Causal Inference

[Pearl et al., 2016]



structural causal models

deterministic: $Y = f(U, \cdots)$

non-deterministic: $Y \sim \mathbb{P}(Y \mid f(U), \cdots)$

how can we make fair predictions?

Counterfactuals

Given an individual:

A_s	A_r	G	L	Y
male	black			

1. Change race attribute

A_s	a'	G	L	Y
male	~~black~~ white			

2. Compute unobserved variables in causal model

A_s	a'	G	L	Y	U
male	~~black~~ white				

3. Recompute observed variables in causal model

A_s	a'	$G_{A_r \leftarrow a'}$	$L_{A_r \leftarrow a'}$	$Y_{A_r \leftarrow a'}$	U
male	~~black~~ white				

Counterfactual Fairness

Definition. *A predictor* $\hat{Y}$ *is* **counterfactually fair** *if given observations* $\mathcal{X} = \mathbf{x}$ *and* $A = a$ *we have that,*

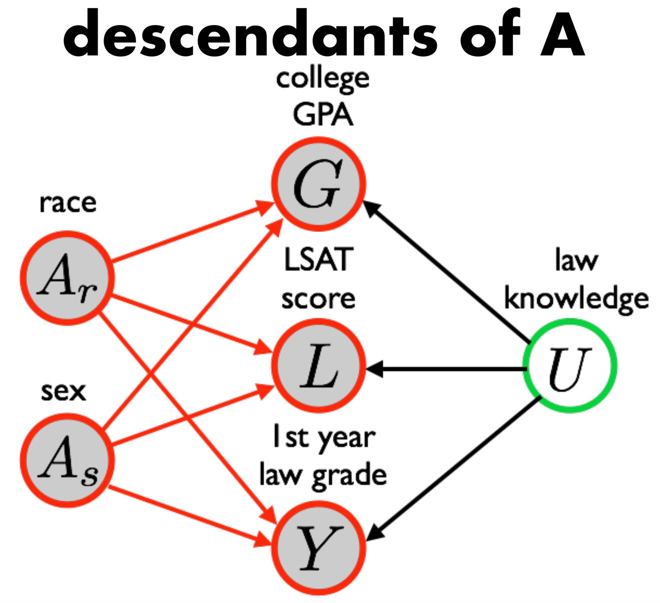
$$\mathbb{P}(\hat{Y}_{A\leftarrow a} = y \mid \mathcal{X} = \mathbf{x}, A = a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'} = y \mid \mathcal{X} = \mathbf{x}, A = a)$$

for all y *and* $a' \neq a$.

Compares **the same individual** with a different version of him/herself

Learning Fair Predictors

Counterfactuals alter descendants of A



Algorithm (non-deterministic models)

Given: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}, a^{(i)})\}_{i=1}^{d}$

a) For each data point $i \in \mathcal{D}$, sample $u_1^{(i)}, \ldots, u_m^{(i)} \sim \mathbb{P}(U \mid \mathbf{x}^{(i)}, a^{(i)})$

b) $\hat{\theta} \leftarrow \arg\min_{\theta} \sum_{i\in\mathcal{D}} \sum_{j=1}^{m} \ell(y^{(i)}, \hat{Y}_\theta(u_j^{(i)}, \mathbf{x}^{(i)}_{\not\succ A}))$

c) Return: $\hat{Y}_{\hat{\theta}}$

features that are not descendants of A

Related Work

[Kilbertus et al., 2017]

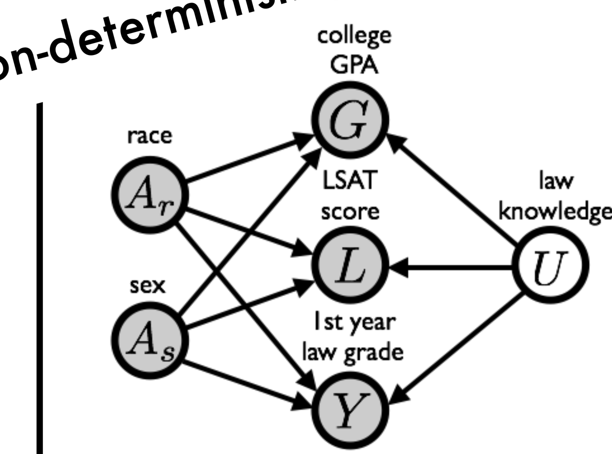
$\mathbb{P}(\hat{Y} = y \mid do(A=a), \mathcal{X} = \mathbf{x}) = \mathbb{P}(\hat{Y} = y \mid do(A=a'), \mathcal{X} = \mathbf{x})$

Compares **different individuals** with the same observed features

Results: US law schools

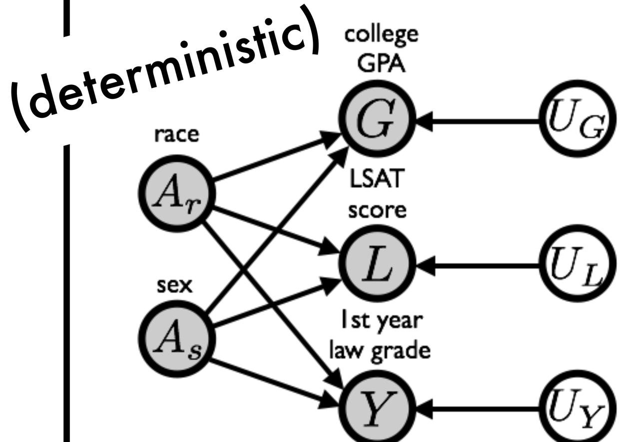
causal models

(non-deterministic)



$G \sim \text{Normal}(b_G + [U, A_r, A_s]\mathbf{w}_G, \sigma_G)$
$L \sim \text{Poisson}(\exp(b_L + [U, A_r, A_s]\mathbf{w}_L))$
$Y \sim \text{Normal}([U, A_r, A_s]\mathbf{w}_Y, 1)$
$U \sim \text{Normal}(0, 1)$

(deterministic)

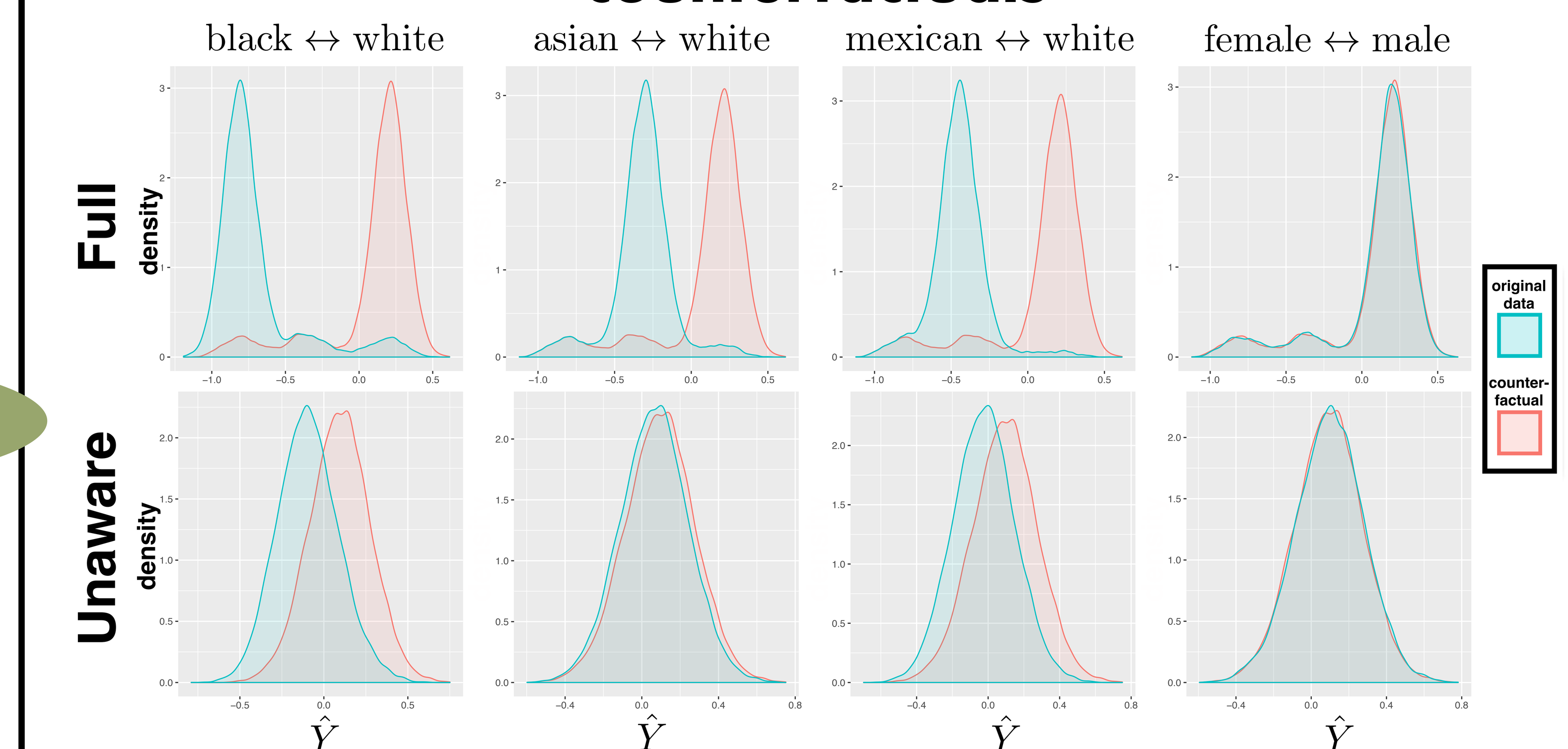


$G = b_G + [A_r, A_s]\mathbf{w}_G + U_G$
$L = b_L + [A_r, A_s]\mathbf{w}_L + U_L$
$Y = b_Y + [A_r, A_s]\mathbf{w}_Y + U_Y$
$U_G \sim \mathbb{P}(U_G)$
$U_L \sim \mathbb{P}(U_L)$
$U_Y \sim \mathbb{P}(U_Y)$

predictive error

	Full	Unaware	C-Fair (Non-Det.)	C-Fair (Det.)
RMSE	0.873	0.894	0.929	0.918

counterfactuals



Results: NYC stop-and-frisk

Arrest rate (data) | Arrest if White (counterfactual) | Arrest if Black Hispanic (counterfactual)