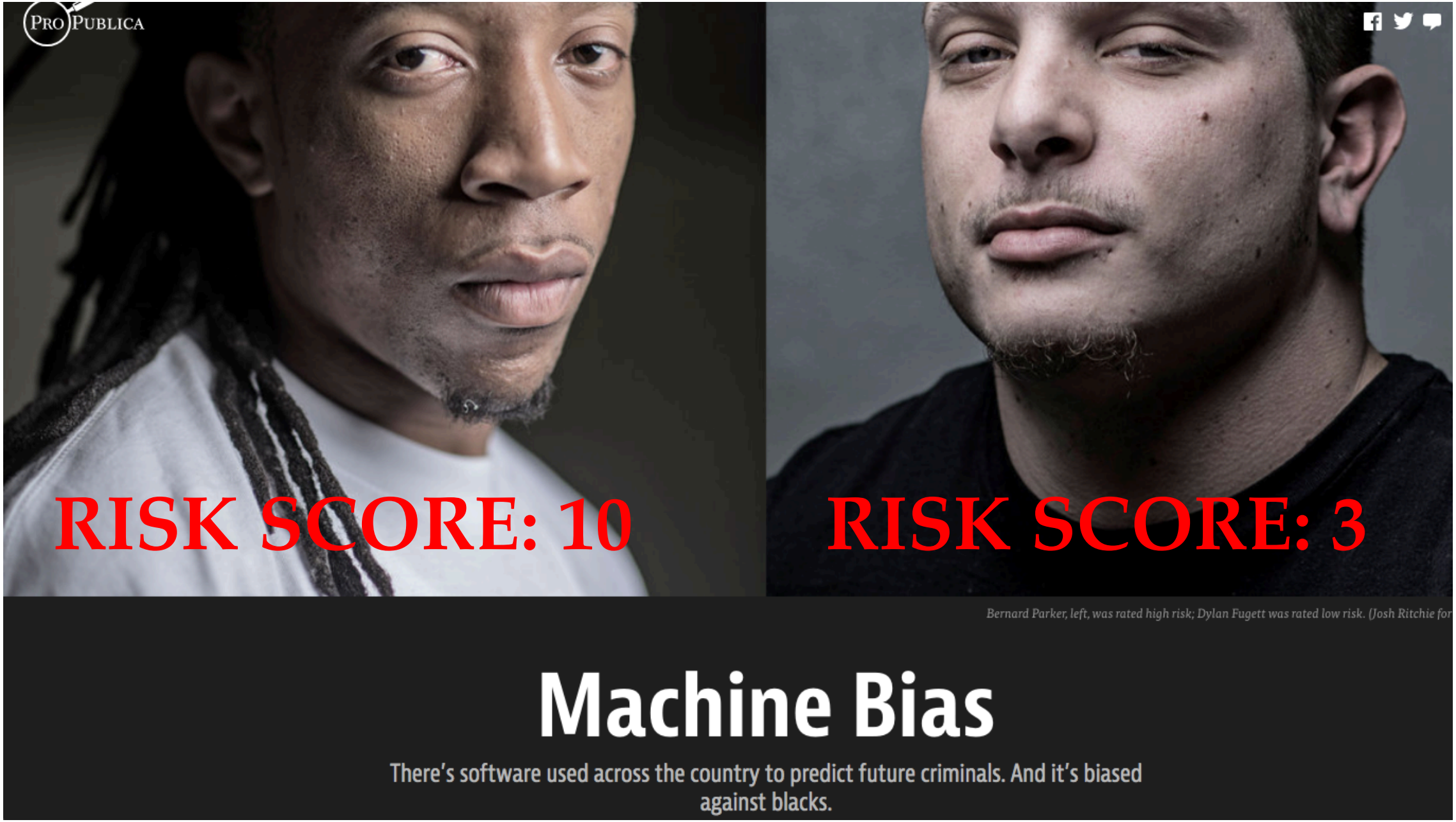
The Bias Report
In Search of Equity in Algorithmic Decision Making

Pedro Saleiro (and Rayid Ghani)
Center for Data Science and Public Policy, University of Chicago
Porto, April 2018

[ProPublica16]



[ProPublica16]

US has 27 Anti-Discrimination Acts

Equal Pay Act of 1963
Civil Rights Acts of1964 – employment
Civil Rights Acts of 1968 – housing
(…)

Types of Discrimination

Disparate Treatment

The process takes into account the protected attribute (e.g. jobs advertised only for age < 50)

Disparate Impact

The process is considered neutral but impacts people from different groups differently (e.g. interviewing more candidates for the job with age < 50)

Algorithmic Decision Making

ADM can be seen as statistical risk assessment to replace or assist human decision making

Goal:

scoring the probability of a binary outcome (yes/no)

given a set of attributes + training data

It is becoming ubiquitous in our lives and has a huge impact

Algorithmic Decision Making

Banking: Is this client at risk of defaulting?

Criminal Justice: Is this suspect likely to re-offend?

HR Management: Is this candidate a good fit?

Healthcare: Is this patient at risk of getting disease X?

Algorithmic Decision Making

ADM-aided systems are usually optimized for a specific global metric (e.g. Precision, AUC)

Often there is a budget constraint, i.e. a fixed limit number of interventions (yes decisions)…

…which represents an adjusted scoring threshold

What is a group?

Computational speaking, a **group** is a set of entities that have in common the **value** of a given **attribute**

Examples:

race = ‘black’
race = ‘hispanic’
race = ‘caucasian’
gender = ’male’
gender = ‘female’
gender = ‘other’

Parity based notion of fairness

Reference Group:

Given an attribute A (e.g. race) we select a reference group (e.g. the historical favored group race = 'white')

Bias metric as disparity:

The ratio between a group metric (e.g. FPR) value of a given group and the reference group

$$FPR_g\ disp = \frac{FPR_{a_i}}{FPR_{a_r}} = \frac{\Pr(\widehat{Y}=1|Y=0,A=a_i)}{\Pr(\widehat{Y}=1|Y=0,A=a_r)}$$

Parity based notion of fairness

This notion requires that all biases (disparities) to be within the range defined by the fairness threshold.

$$\tau \leq DisparityMeasure_{group_i} \leq \frac{1}{\tau}$$

Example: If fairness threshold is 0.8, the fairness range is between 80% and 125% of the group metric value of the selected reference group.

No “Fairness through Unawareness”

Remove protected attributes?

Well, other features subsume the protected attributes.

Example: Easy to predict gender based on Facebook likes.

No “Fairness through Statistical Parity”

Decision to be independent from the protected attribute?

Example: Accept equal number of students from every race in a given master program.

Does not ensures “supervised fairness”, as it is possible to have different false positive/negative parities across groups.

Cripples the overall utility metric (e.g. A correlated with Y)

Fairness Tradeoffs

If the base rate (prevalence) is different between groups and the classifier is non-trivial (TPR>0) and imperfect (FPR > 0). Then, either:

Precision Parity Fails (no calibration is possible)

FPR and FNR will be disparate (no equalized odds)

[Kleinberg16, Chouldechova17]

Aequitas

How can you use Aequitas?

Web Audit Tool

Try our Audit Tool to generate a Bias Report

1. Upload Data (or use pre-loaded sample data)
2. Configure (bias metrics of interest and reference groups)
3. Generate the Bias Report

Try it out! ›

Python Library

Use our python code library to generate bias and fairness metrics on your data and predictions.

Python Code ›

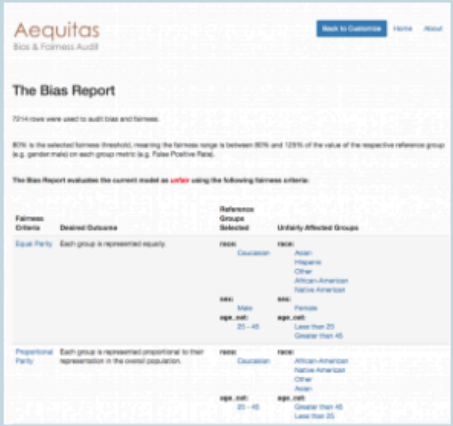
Command Line Tool

Use our command line tool to generate a report using your own data and predictions.

Aequitas

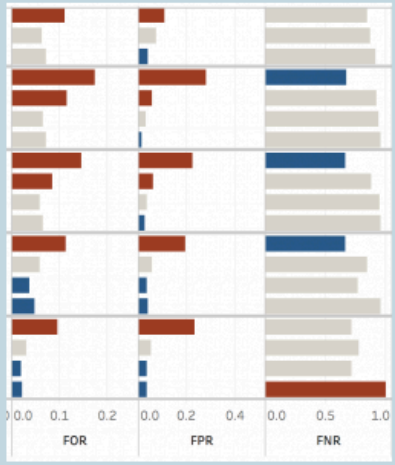
What does Aequitas produce?

Bias Report

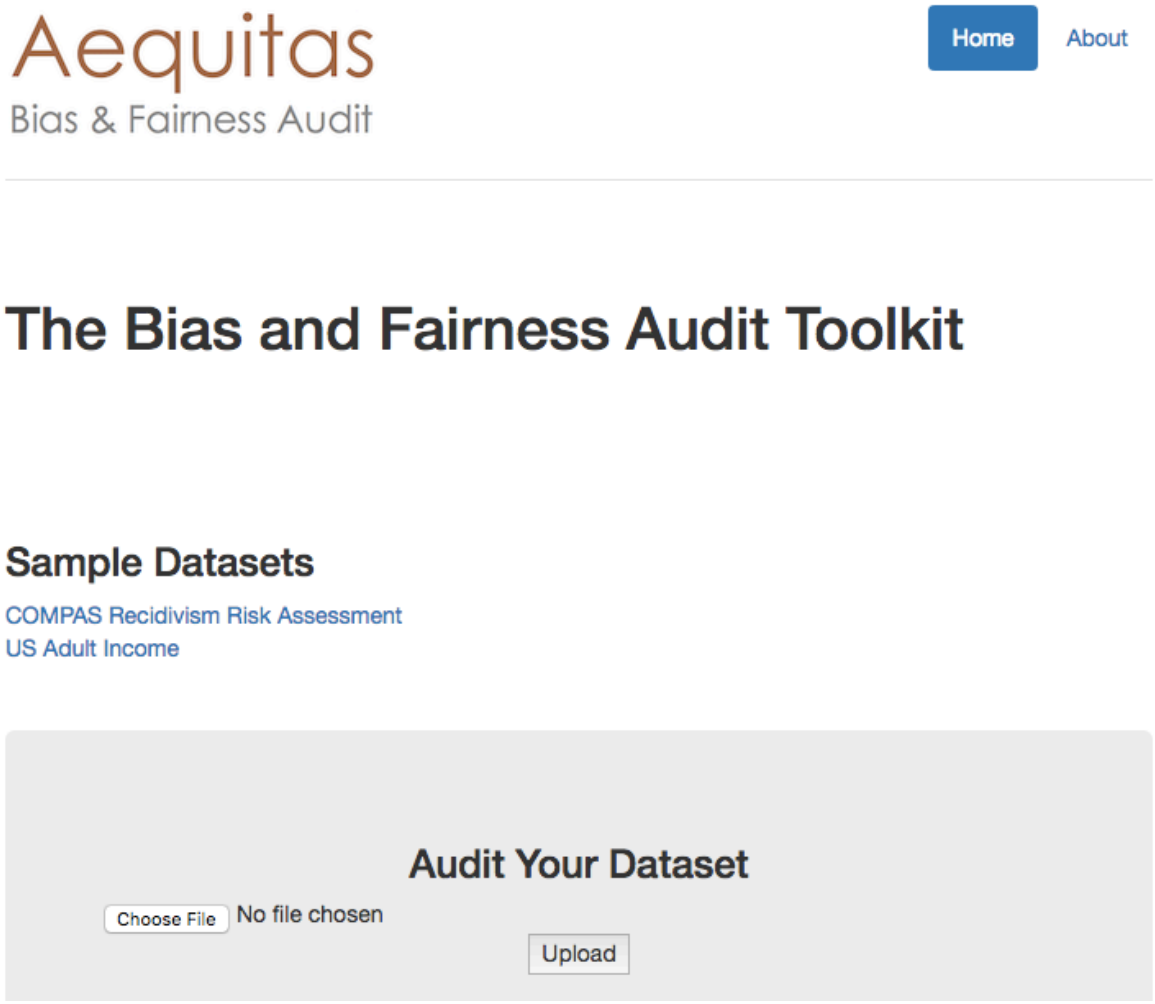


Detailed Fairness and Bias Statistics

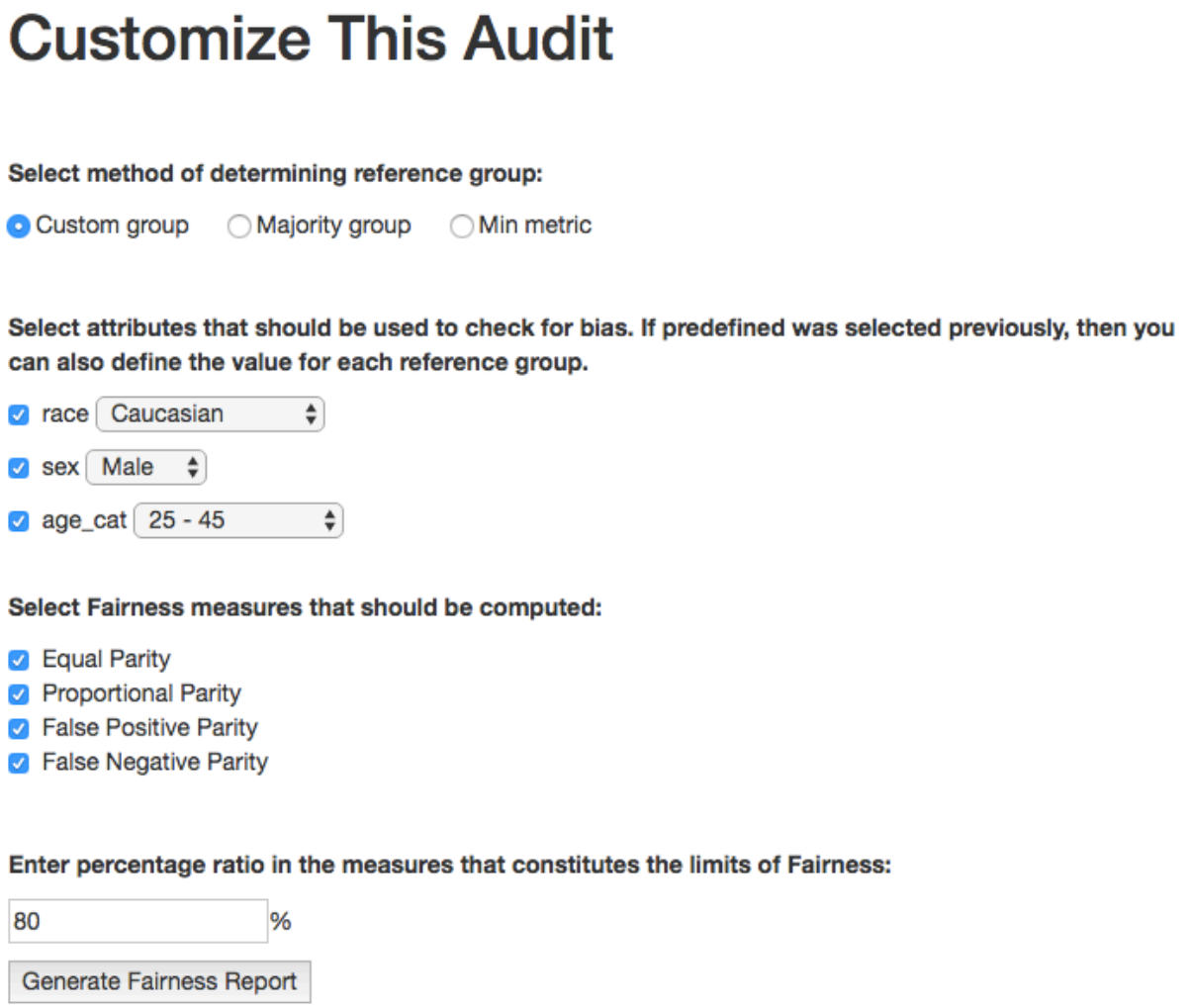
Interactive Bias Dashboard



Aequitas


Aequitas
Bias & Fairness Audit
Home
About
The Bias and Fairness Audit Toolkit
Sample Datasets
COMPAS Recidivism Risk Assessment
US Adult Income
Audit Your Dataset
Choose File
No file chosen
Upload

Aequitas


Customize This Audit
Select method of determining reference group:
Custom group
Majority group
Min metric
Select attributes that should be used to check for bias. If predefined was selected previously, then you can also define the value for each reference group.
race
Caucasian
sex
Male
age_cat
25 - 45
Select Fairness measures that should be computed:
Equal Parity
Proportional Parity
False Positive Parity
False Negative Parity
Enter percentage ratio in the measures that constitutes the limits of Fairness:
80
%
Generate Fairness Report

Aequitas

The Bias Report

7214 rows were used to audit bias and fairness.

80% is the selected fairness threshold, meaning the fairness range is between 80% and 125% of the value of the respective reference group (e.g. gender:male) on each group metric (e.g. False Positive Rate).

The Bias Report evaluates the current model as unfair using the following fairness criteria:

Fairness Criteria	Desired Outcome	Reference Groups Selected	Unfairly Affected Groups
Equal Parity	Each group is represented equally.	**race:** Caucasian **sex:** Male **age_cat:** 25 - 45	**race:** Asian Hispanic Other African-American Native American **sex:** Female **age_cat:** Less than 25 Greater than 45

Aequitas

False Positive Parity

False Positive Parity is concerned with Type I errors (False Positives). In cases of punitive interventions on the selected set it is important to not have disparate Type I errors across groups. Aequitas audits both False Positive Rate (FP/Negative Labels of each group) and False Discovery Rates (FP/Selected Set Size).

False Positive Rate

What is it?	When should I care about False Positive Parity?	Unfairly Affected Groups
This criteria considers an attribute to have False Positive parity if every group has the same False Positive Error Rate. For example, if race has false positive parity, it implies that all three races have the same False Positive Error Rate.	If your desired outcome is to make false positive errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is punitive and has risk of adverse consequences for the selected set. Using this criteria allows you to make sure that you are not making mistakes about any single group disproportionately.	**race:** Native American: **160**% of the false positive rate of the reference group "Caucasian", corresponding to a difference of 0.38 vs 0.23. Other: **63**% of the false positive rate of the reference group "Caucasian", corresponding to a difference of 0.15 vs 0.23. Asian: **37**% of the false positive rate of the reference group "Caucasian", corresponding to a difference of 0.09 vs 0.23. African-American: **191**% of the false positive rate of the reference group "Caucasian", corresponding to a difference of 0.45 vs 0.23. **age_cat:** Greater than 45: **50**% of the false positive rate of the reference group "25 - 45", corresponding to a difference of 0.17 vs 0.33. Less than 25: **162**% of the false positive rate of the reference group "25 - 45", corresponding to a difference of 0.54 vs 0.33.

Aequitas

race

Attribute Value	Equal Parity	Impact Parity	FDR Parity	FPR Parity	FOR Parity	FNR Parity
African-American	Unfair	Unfair	Fair	Unfair	Fair	Unfair
Asian	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair
Caucasian	Ref	Ref	Ref	Ref	Ref	Ref
Hispanic	Unfair	Fair	Fair	Fair	Fair	Fair
Native American	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair
Other	Unfair	Unfair	Fair	Unfair	Fair	Unfair

Aequitas

race

Attribute Value	PPR Disparity	PPREV Disparity	FDR Disparity	FPR Disparity	FOR Disparity	FNR Disparity
African-American	2.55	1.69	0.91	1.91	1.21	0.59
Asian	0.01	0.72	0.61	0.37	0.43	0.7
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.22	0.86	1.12	0.92	1.0	1.17
Native American	0.01	1.92	0.61	1.6	0.58	0.21
Other	0.09	0.6	1.12	0.63	1.05	1.42

Aequitas

race

Attribute Value	Group Size Ratio	PPR	PPREV	FDR	FPR	FOR	FNR
African-American	0.51	0.66	0.59	0.37	0.45	0.35	0.28
Asian	0	0.0	0.25	0.25	0.09	0.12	0.33
Caucasian	0.34	0.26	0.35	0.41	0.23	0.29	0.48
Hispanic	0.09	0.06	0.3	0.46	0.21	0.29	0.56
Native American	0	0.0	0.67	0.25	0.38	0.17	0.1
Other	0.05	0.02	0.21	0.46	0.15	0.3	0.68

Real-world Public Policy Projects

Criminal Justice

Public Safety and Policing

Public Health

Housing Safety

Criminal Justice

Goal: use historical data about individuals, predict their probability of recidivism in the next 6 months, and match the 150 highest risk individuals with tailored, preventative interventions.

Data: criminal justice data for 1.5 Million individuals over the past 10 years, focusing on the 400K individuals who had repeated interactions with the justice system.

Performance Metric: Precision @ 150 absolute

Public Safety and Policing

Goal: Given the set of all active officers at a given date and all data collected by a police department prior to that date, predict which officers will have an adverse interaction in the next year.

Data: The data for this work consists of almost all employee information and event records collected by the Police Department to manage its day-to-day operations.

Performance Metric: Precision @ 10%

Public Health

Goal: The goal of this work was to build a point-of-service machine learning system that, at the time of a clinical visit, assesses the HIV patient's risk of not returning for continued treatment, as well as the associated risk factors.

Data: Electronic health records of all the patients receiving care from an HIV Clinic. The patient records span 8 years from 2008 - 2016 and include approximately 1,600 patients.

Performance Metric: Precision @ 10%

Housing Safety

Goal: Increase the safety and well-being of residents living in many rental properties by prioritizing inspection in rental properties more likely to have serious health and safety violations (every quarter 300 houses)

Data: The data used consisted of all housing inspections that were done by the city, outcomes of those inspections, cases that were initiated as a result, and the outcomes of those cases.

Performance Metric: Precision @ 300

Real-world Public Policy Projects

The ADM systems we developed for these problems do indeed have bias...

... but in most cases the alternatives being used by policymakers today are much more biased.

The ADM systems tend to be more accurate, and either equally or less biased, in effect improving equity and fairness of the policy.

Fairness in Algorithm Decision Making is not an absolute statistical concept!

It depends on the application scenario and respective social and ethical impacts!

Thank You!

Questions?

saleiro at uchicago.edu

References

[USfed] U.s. federal legislation. (a) equal credit opportunity act, 1974; (b) fair housing act, 1968; (c) employment act, 1967; (d) equal pay act, 1963; (e) pregnancy discrimination act, 1978; (f) civil right act, 1964, 1991.

[Propublica16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, May, 23, 2016.

[Hardt16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pages 3315–3323, 2016.

[Kleinberg16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.

[Chouldechova17]