ML Bias Bootcamp

INFO 251

Dr. Nick Merrill, UC Berkeley
cltc.berkeley.edu

About me

Nick Merrill

I direct the Daylight Lab at the UC Berkeley Center for Long-Term Cybersecurity.

Our thesis: Cybersecurity is hard to practice in part because it's hard to understand. **Our mission**: To help people understand the cybersecurity issues that matter to them.

About this bootcamp

MLFailures

Goal: To make machine learning bias easier to **identify and ameliorate**.

Impact: This **open-access** bootcamp is taught to students, policymakers, and engineers around the world every year.

Learn more at **https://daylight.berkeley.edu/mlfailures**.

Part 1

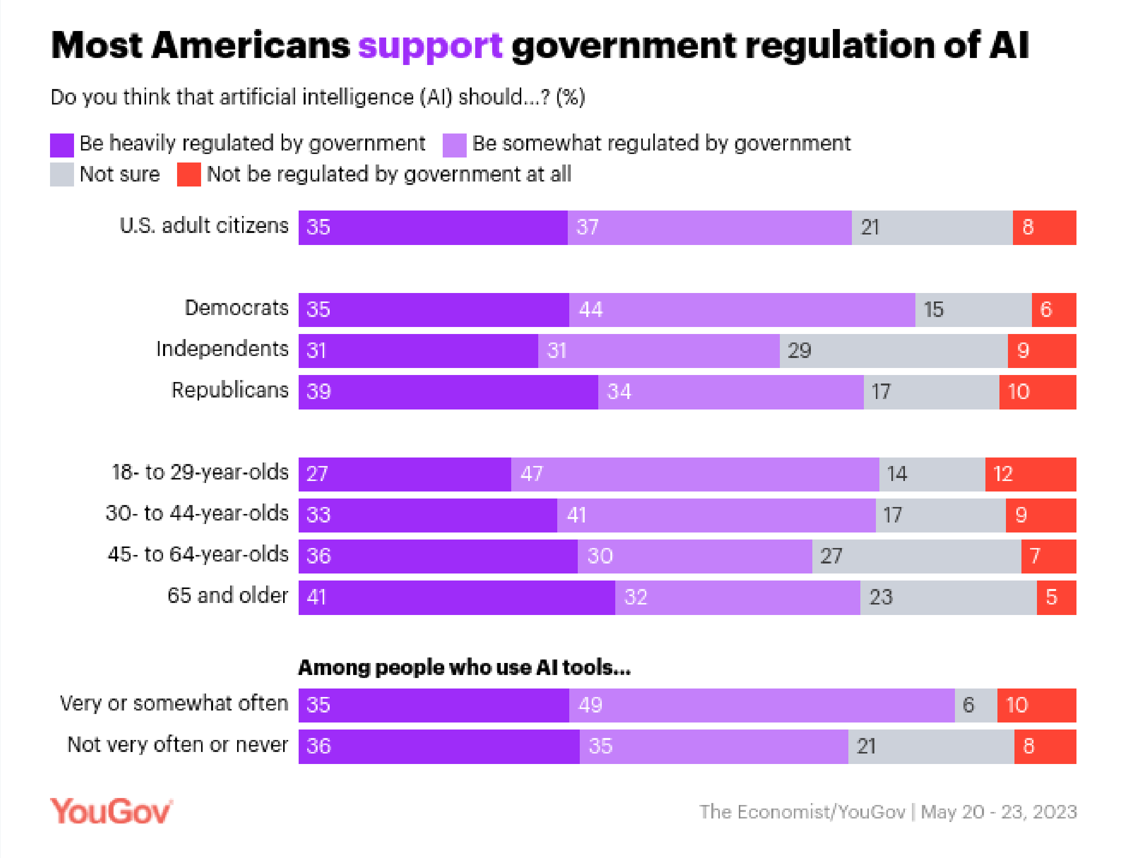
What is bias?

EDITORS' PICK

Exclusive: DHS Used Clearview AI Facial Recognition In Thousands Of Child Exploitation Cold Cases

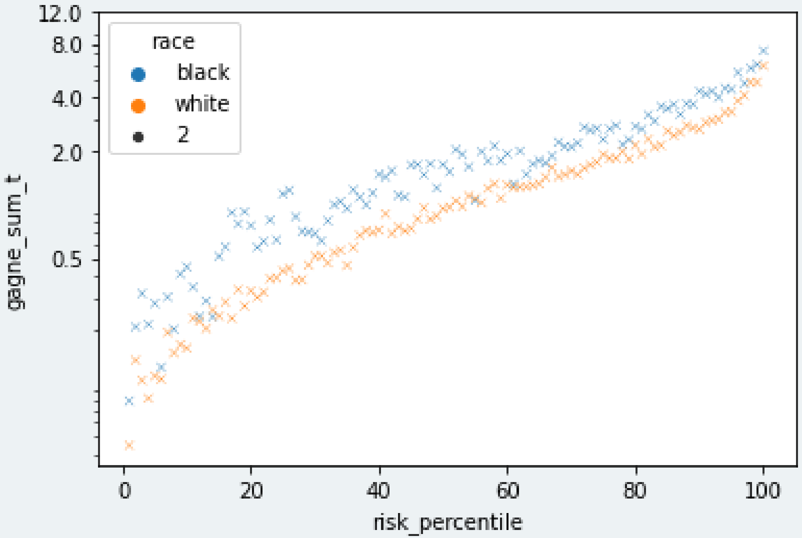
Eight Months Pregnant and Arrested After False Facial Recognition Match

Porcha Woodruff thought the police who showed up at her door to arrest her for carjacking were joking. She is the first woman known to be wrongfully accused as a result of facial recognition technology.

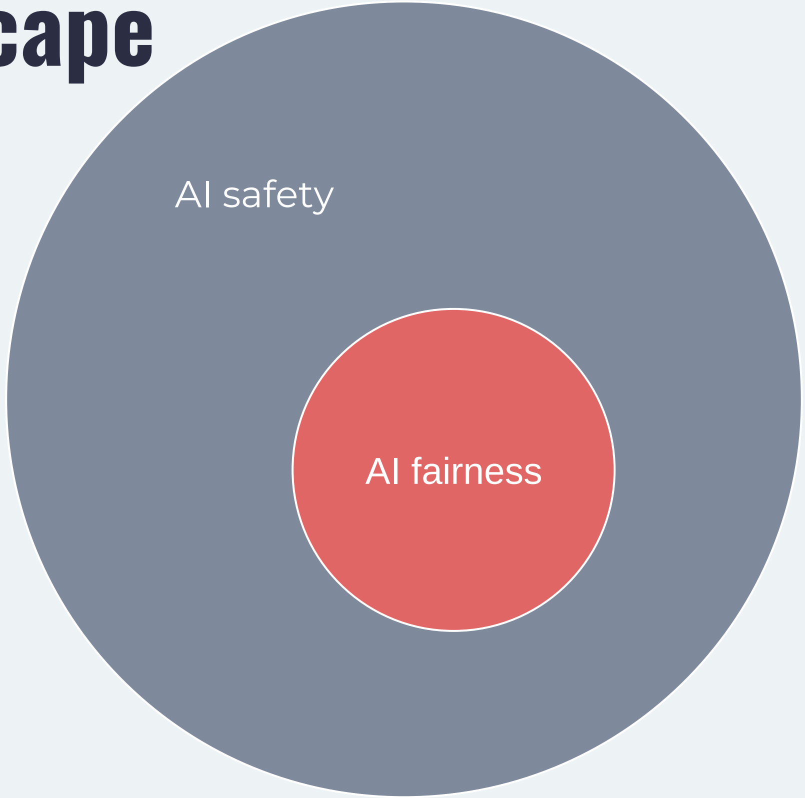

Most Americans support government regulation of AI
Do you think that artificial intelligence (AI) should...? (%)
Be heavily regulated by government
Be somewhat regulated by government
Not sure
Not be regulated by government at all
U.S. adult citizens 35 37 21 8
Democrats 35 44 15 6
Independents 31 31 29 9
Republicans 39 34 17 10
18- to 29-year-olds 27 47 14 12
30- to 44-year-olds 33 41 17 9
45- to 64-year-olds 36 30 27 7
65 and older 41 32 23 5
Among people who use AI tools...
Very or somewhat often 35 49 6 10
Not very often or never 36 35 21 8
YouGov
The Economist/YouGov | May 20 - 23, 2023

... in Healthcare,

an ML Failure leads to... worse medical care for Black patients than for white patients.



The landscape



Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Defining ML bias

Hardt (2017) defines a biased model as one that exhibits **systematically adverse discrimination** based on "**socially salient qualities** that have **served as the basis for unjustified and systematically adverse treatment** in the past."

Sensitive features

Sensitive features are features that represent “socially salient” characteristics:

- Family status
- Sexual orientation
- Veteran status
- Disability status
- ...

Is it a sensitive feature?

Gender	Yes
Race	Yes
School district	No! But it's probably *correlated* with a sensitive feature or two...
Number of seconds the cursor spent hovering over the search bar	No! But it's *may be correlated* with a sensitive feature or two...

Types of Errors

Imbalanced Data Sets

→ Amazon's hiring model
→ Google's CV model

Removing Sensitive Features

→ If you've removed race, how will you know if it exhibits racial bias?

Lurking Bias in Features

→ e.g., Credit score is predictive of race

Poorly Framed Problems

→ Predicting if someone is a criminal based on their face

Questions?

See for yourself

Let’s walk through an interactive example that explores fairness in the context of **healthcare data**.

Background

Risk-scoring in hospitals

- Nurses need to triage patients accurately
  - Nurses want to triage according to *medical risk*
  - “High-risk care management” – extra care that may make a life-or-death difference.
- But… Nurses are extremely busy, overworked, and tired.
- Algorithms (cl)aim to
  - Automate risk scoring, making nurses’ lives easier
  - Lower cost of care
  - Improve patient outcomes
- This algorithm is *used pervasively*
  - Applied to ~200MM people in the US every year
- This algorithm is produced by private companies

Background

Features the algorithm uses:

- Demographics (age, sex)
- Insurance type
- Diagnosis & procedure codes
- Medication
- Medical costs

Any **red flags**?

Background

Features the algorithm uses:

- Demographics (age, sex)
- Insurance type
- Diagnosis & procedure codes
- Medication
- Medical costs

Features it doesn't use:

- **Race**

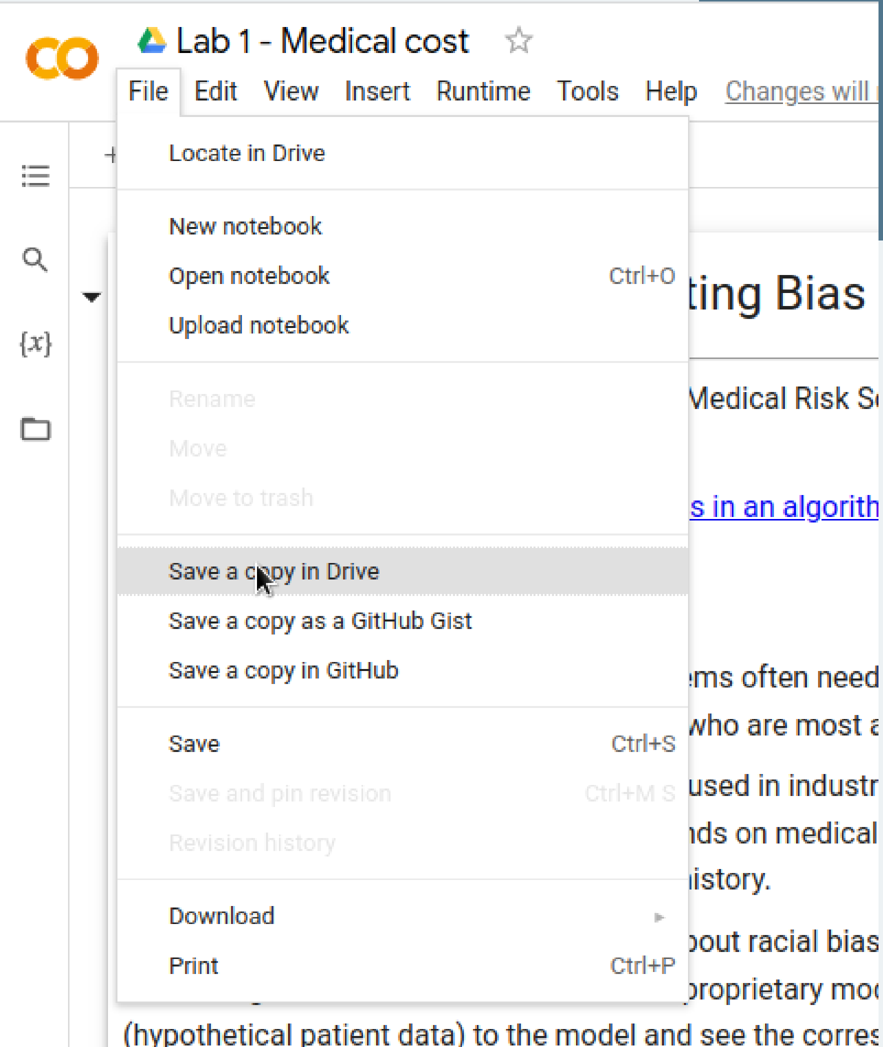
Obermeyer et al added race back into this dataset **(manually; their work was published in Nature)**

They asked: **do we see a difference in treatment based on race**?

Let's get started!

https://tinyurl.com/2p8bhpuk

File > Save a copy in Drive



What went wrong?

Features the algorithm used:

- Demographics (age, sex)
- Insurance type
- Diagnosis & procedure codes
- Medication
- Medical costs

Which feature was a cause of the bias?

What went wrong?

Features the algorithm used:

- Demographics (age, sex)
- Insurance type
- Diagnosis & procedure codes
- Medication
- **Medical costs**

Which feature was a cause of the bias?

Medical cost acts as a proxy of need… but also of **access to care**.

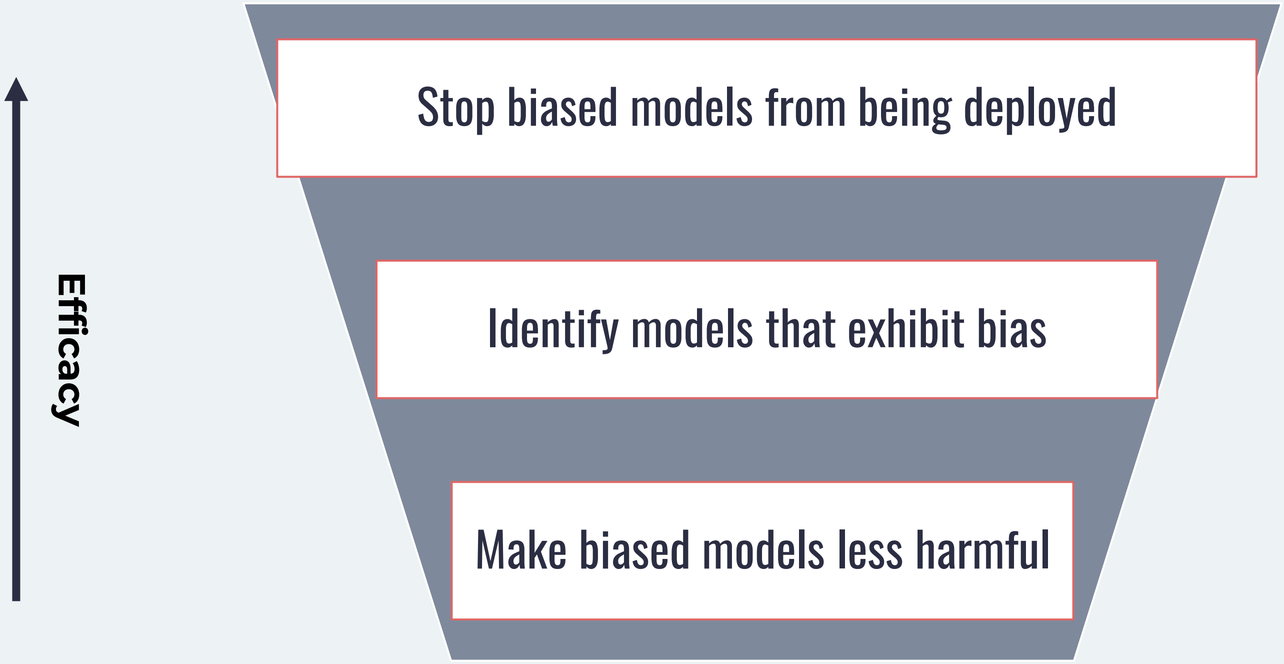
Consider:

- Medical insurance
- Doctors' attitudes toward patients
- Proximity of quality hospitals
- Etc…

Part 2

What can we do about ML bias?

The “funnel” of prevention



The “funnel” of prevention

Efficacy ↑

- Stop biased models from being deployed — Internal audits
- Identify models that exhibit bias — External audits
- Make biased models less harmful — ***Last resort***: Engineer patches to models

Identifying bias

There are technical strategies for auditing bias.

Disparate impact

The issues salient to ML bias are *not* completely *de novo*. There's legal precedent for evaluating biased treatment and impact

- **80% rule**
  - 1978 Uniform Guidelines on Employee Selection Procedure
  - If the selection rate for a certain group is less than 80% of rate for group with highest selection rate, there is disparate impact. (Biddle, 2006).
    - Example: A landlord accepts 60% of applications from white tenants, but only 40% of applications from Black tenants.
  - Even if there is no evidence of intentional disparate treatment, the disparate impact is *still* a violation.

Strategies to Identify Bias

DISPARATE IMPACT

STATISTICAL PARITY

EQUAL OPPORTUNITY

AVERAGE ODDS

Each group should have an equal opportunity of achieving the favorable outcome.
We calculate the ratio of rate of favorable outcome for unprivileged group compared to that of privileged group.

The ideal value is 1.
A value < 1 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	0.55	Sex	Male	0.29
Recidivism (Compas)	Race	White	0.75	Sex	Female	0.59

Law and this metric are not the same (Watkins et al, 2022)

Strategies to Identify Bias

DISPARATE IMPACT
STATISTICAL PARITY
EQUAL OPPORTUNITY
AVERAGE ODDS

Statistical Parity:
Demographics of those receiving any classification should be the same as demographics of the underlying population.

Equal Opportunity / Average Odds
Each group should be classified (in)correctly at the same rate.

Look to the appendix to learn about these other strategies for identifying bias.

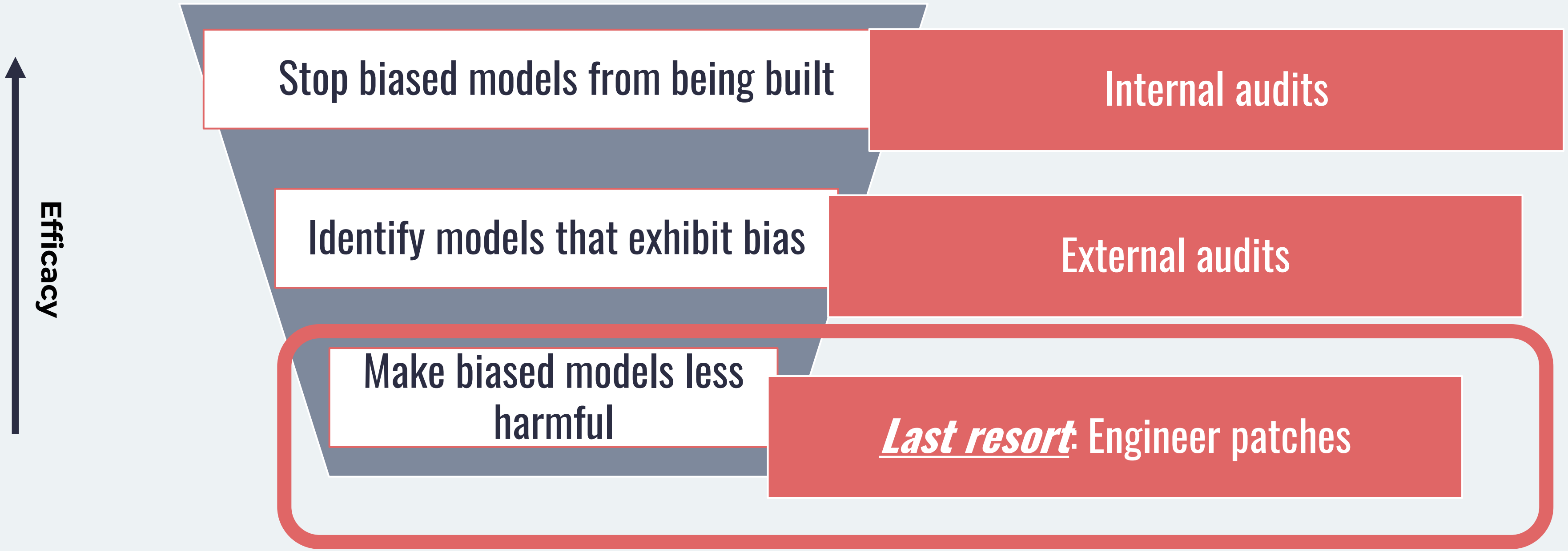
Takeaway: audits and policy can harmonize

- **Existing policy & legislation could be applied to ML**… *if* we have access to **sensitive features**.
- Disparate impact is an example of a legal theory ML audits can speak to.
- New policies can harmonize with existing identification strategies.

Ameliorating bias

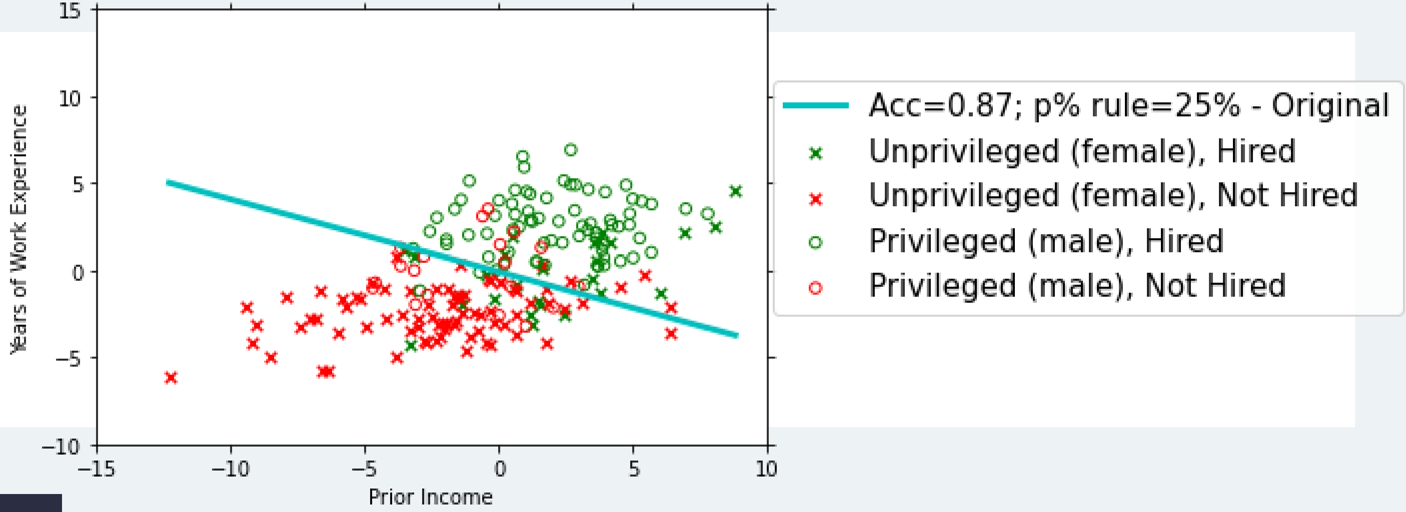
While there is no way to “fix” bias, **there are methods for making bias less harmful**.

The “funnel” of prevention

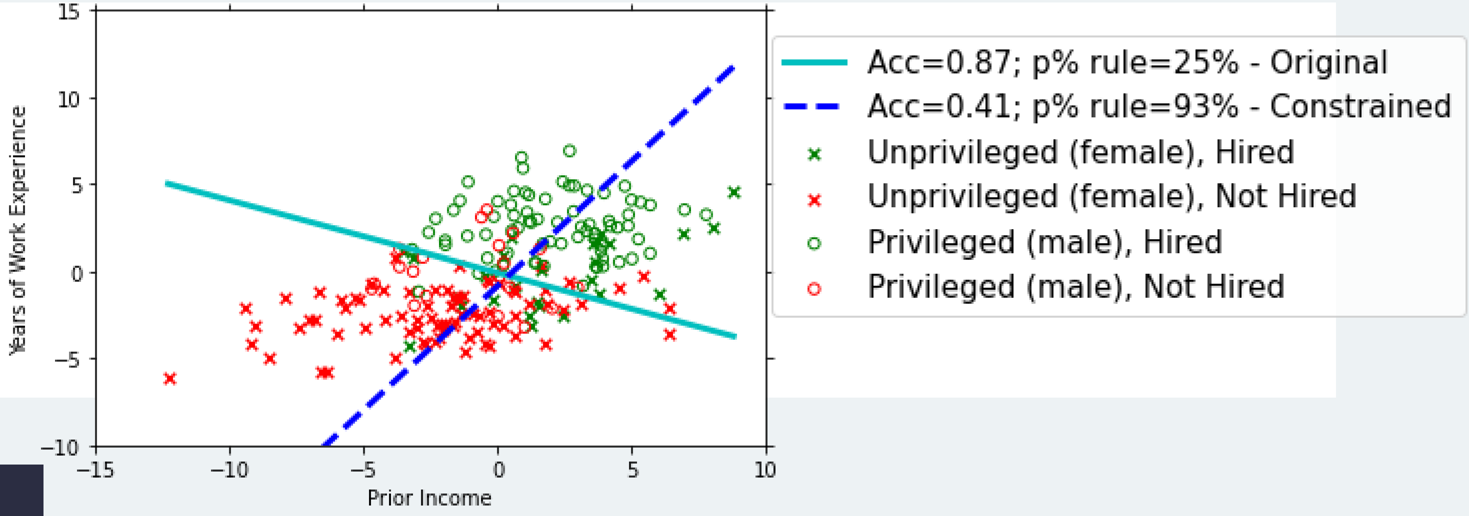


Gender bias in hiring

Gender bias in hiring



Gender bias in hiring



Strategies to Mitigate Bias

CONSTRAINTS
REWEIGHTING
OPTIMIZED PRE-PROCESSING
ADVERSARIAL DEBIASING
REJECT-OPTION-BASED CLASSIFICATION

Fairness constraints allow us to specify a tradeoff between a classifier's "fairness" and its accuracy.
Sometimes, our dataset is badly biased. For example, a dataset of past hiring decisions may embed a bias against women. In this case, an "accurate" classifier would be unfair - perhaps illegally so.

To correct for this, we can set a fairness constraint (e.g., a minimum disparate impact score).
With this constraint, the classifier will be as accurate as possible while exceeding the minimum disparate impact score.

If you are interested, you can see fairness constraints at work in this lab, which focuses on gender bias in a hiring algorithm.

Strategies to Mitigate Bias

- **REWEIGHTING**
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

Strategies to Mitigate Bias

- REWEIGHTING
- **OPTIMIZED PRE-PROCESSING**
- ADVERSARIAL DEBIASING
- REJECT-OPTION-BASED CLASSIFICATION

Learns a probabilistic transformation that can modify the features and the labels in the training data.

Strategies to Mitigate Bias

- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- **ADVERSARIAL DEBIASING**
- REJECT-OPTION-BASED CLASSIFICATION

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.

Since the predictions cannot carry any group discrimination information that the adversary can exploit, the classifier must be fair (right?).

Strategies to Mitigate Bias

- REWEIGHTING
- OPTIMIZED PRE-PROCESSING
- ADVERSARIAL DEBIASING
- **REJECT-OPTION-BASED CLASSIFICATION**

Changes predictions from a classifier to make them fairer.

Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

Everything you need to know about bias

- ML bias is "*sociotechnical*"—it is not **caused** by technical problems alone, and **cannot be "solved"** by technical solutions alone

- Technical approaches can help **identify** and (*to a point*) **ameliorate** ML bias.

- Identifying bias requires audits, and **audits require access to sensitive features**.

Over the next 3-5 years, decisions about AI will be driven by…

1. Technical capacities of AI systems
2. Economic constraints
   - Cost of AI vs. cost of labor
3. **Policy constraints**
   - Import restrictions on AI / de-globalization
   - You must not use AI for...
   - You *must* use AI for…
   - **All AI solutions must…** ⟵ ***We are here.***

Reinforcement learning is going to change everything!

Thank you!

Reach out any time:

- ffff@berkeley.edu
- https://else.how

XX

APPENDIX

Terminology

Privileged Group

We expect this group to get the favorable outcome **more often** than they should.

Unprivileged Group

We expect this group to get the favorable outcome **less often** than they should.

		Privileged Group	Unprivileged Group
Adult Census Income	Race	White	Non-White
	Sex	Male	Non-Male
Recidivism (Compas)	Race	White	Non-White
	Sex	Female	Male

A note on the formalism

Privilege

Systemic inequality and disparate treatment resulting from societal differences in which demographic groups hold power

"Privileged group"

A mathematical formalism describing a group (any group) that gets a favorable outcome (any outcome) more often than they "should"

MIND THE GAP!

→ The formalism of "privileged groups" can help us *understand* privilege - but the two are ***not*** the same.

Thanks to Lauren Chambers

Strategies to Identify Bias

- DISPARATE IMPACT
- **STATISTICAL PARITY**
- EQUAL OPPORTUNITY
- AVERAGE ODDS

Demographics of those receiving any classification should be the same as demographics of the underlying population. We take the difference of rate of favorable outcomes by rate of favorable outcomes by unprivileged group.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.18	Sex	Male	-0.33
Recidivism (Compas)	Race	White	-0.18	Sex	Female	-0.36

Further reading: On the moral justification of statistical parity.

Strategies to Identify Bias

- DISPARATE IMPACT
- STATISTICAL PARITY
- **EQUAL OPPORTUNITY**
- AVERAGE ODDS

Each group should be ‘equally’ incorrectly classified.
We take the difference of true positive rates between unprivileged and privileged groups.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.06	Sex	Male	-0.14
Recidivism (Compas)	Race	White	-0.12	Sex	Female	-0.30

Where is there the **most bias** in these 2 datasets?

Strategies to Identify Bias

- DISPARATE IMPACT
- STATISTICAL PARITY
- EQUAL OPPORTUNITY
- **AVERAGE ODDS**

Each group should be 'equally' incorrectly classified.
We take the average difference of false positive rate and true positive rate between unprivileged and privileged groups.

The ideal value is 0.
A value < 0 implies there is benefit toward the privileged group.

	Trait 1	Privileged	Ratio	Trait 2	Privileged	Ratio
Adult Income	Race	White	-0.09	Sex	Male	-0.19
Recidivism (Compas)	Race	White	-0.16	Sex	Female	-0.35

Where is there the **most bias** in these 2 datasets?