Responsible Data Science

Fairness and Causality

Prof. Elisha Cohen

Center for Data Science
New York University

data RESPONSIBLY

NYU

Fairness and causality

1. (Im)possibility of Fairness
2. Fairness measures
3. Causal models
4. Causal models as a framework for fairness

Reading: Fairness and causality

Big Data
Volume 5 Number 2, 2017

DOI: 10.1089/big.2016.0047

ORIGINAL ARTICLE

Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments

Alexandra Chouldechova*

Abstract

Recidivism prediction instruments (RPIs) provide decision-makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. Although such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This article discusses several fairness criteria that have recently been applied to assess the fairness of RPIs. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when an RPI fails to satisfy the criterion of error rate balance.

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

r/ai

on the (im)possibility of fairness

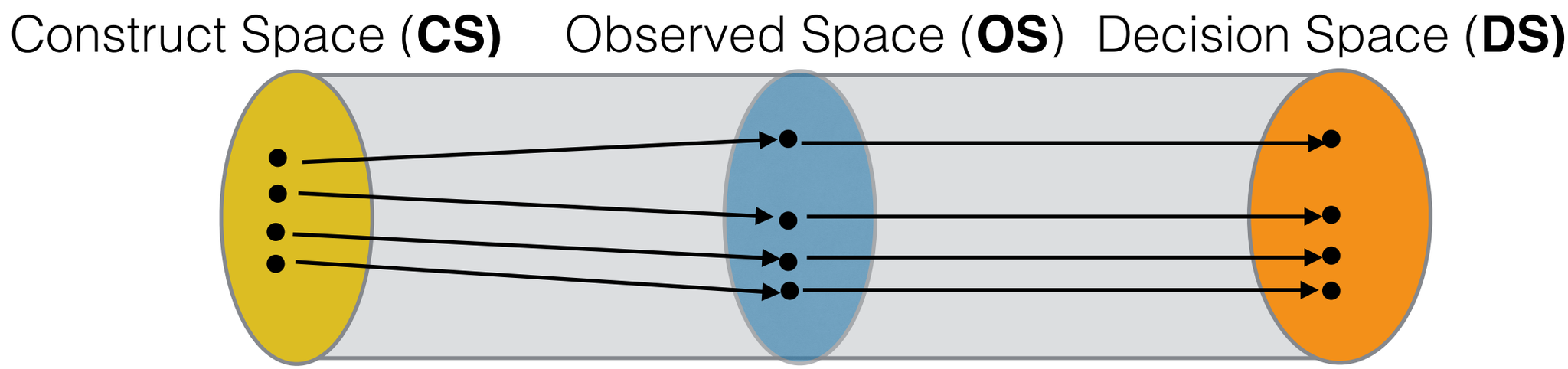
r/ai

On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Goal: tease out the difference between *beliefs* and *mechanisms* that logically follow from those beliefs.

Main insight: To study algorithmic fairness is to study the interactions between different spaces that make up the decision pipeline for a task



On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Construct Space	Observed Space	Decision Space
intelligence	SAT score	performance in college
grit	high-school GPA	
propensity to commit crime	family history	recidivism
risk-averseness	age	

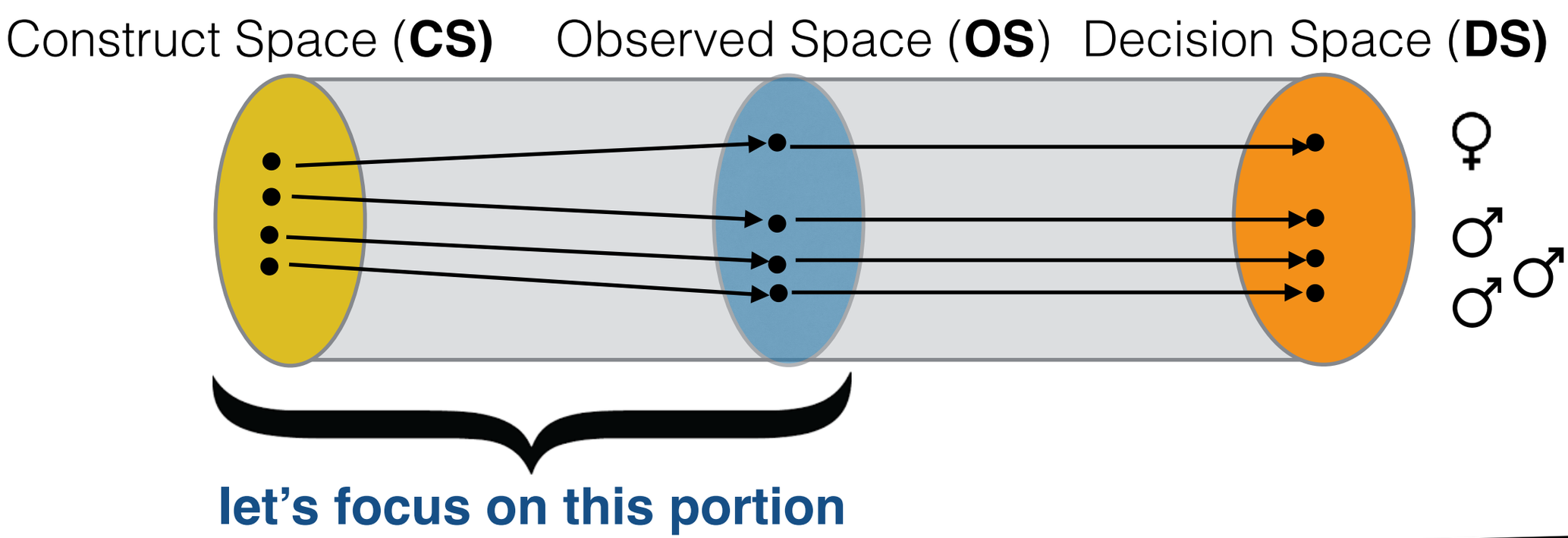
define fairness through properties of mappings

Fairness through mappings

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

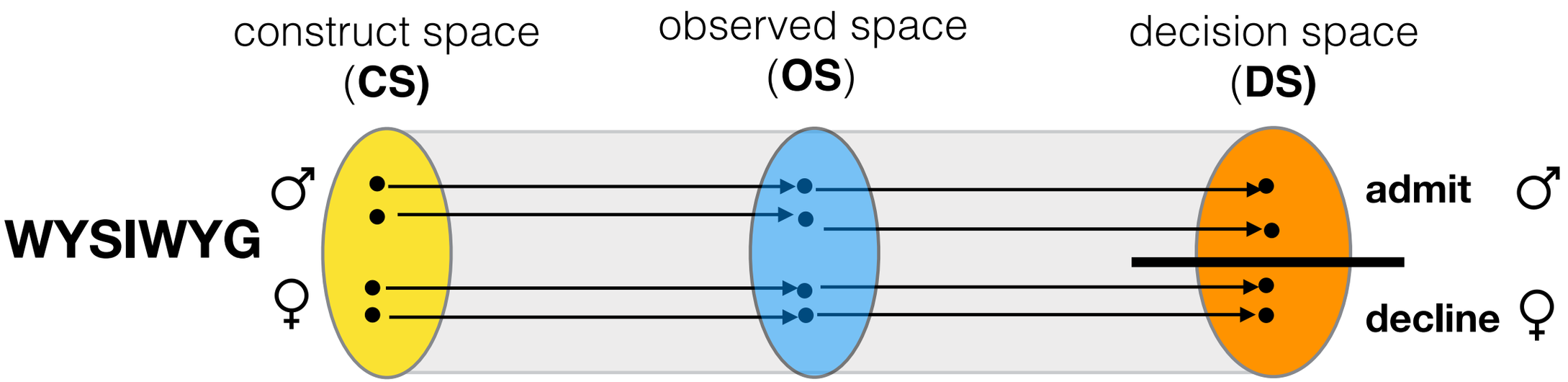
Fairness: a mapping from **CS** to **DS** is **(ε, ε')**-fair if two objects that are no further than **ε** in **CS** map to objects that are no further than **ε'** in **DS**.

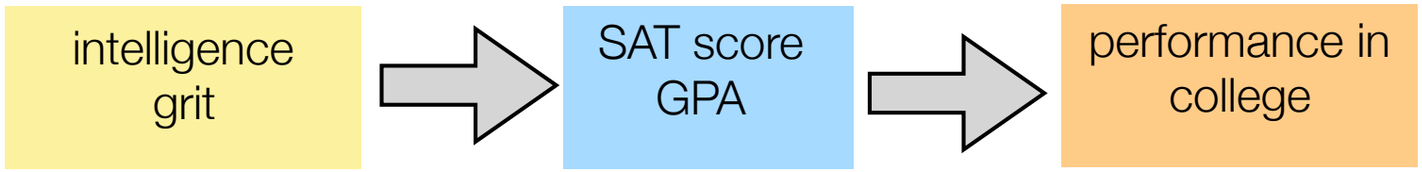
$$f : CS \to DS \qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



WYSWYG

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

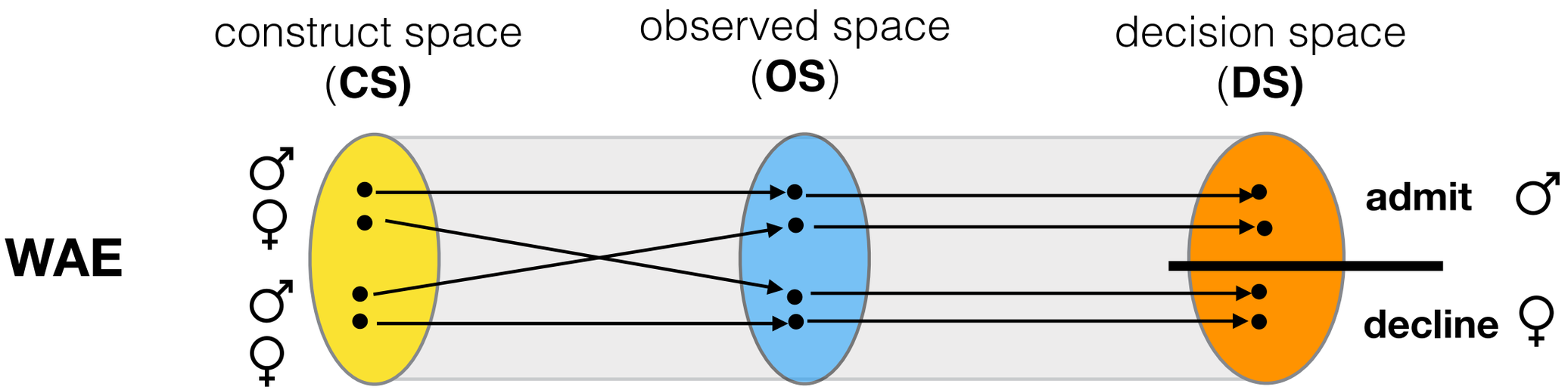


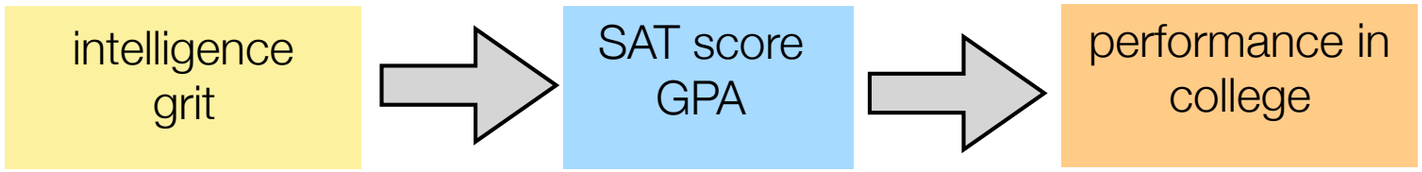


What you see is what you get (**WYSIWYG**): there exists a mapping from **CS** to **OS** that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

WAE

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]





We are all equal (**WAE**): the mapping from **CS** to **OS** introduces **structural bias** - there is a distortion that aligns with the group structure of **CS**. **This is the group fairness world view.**

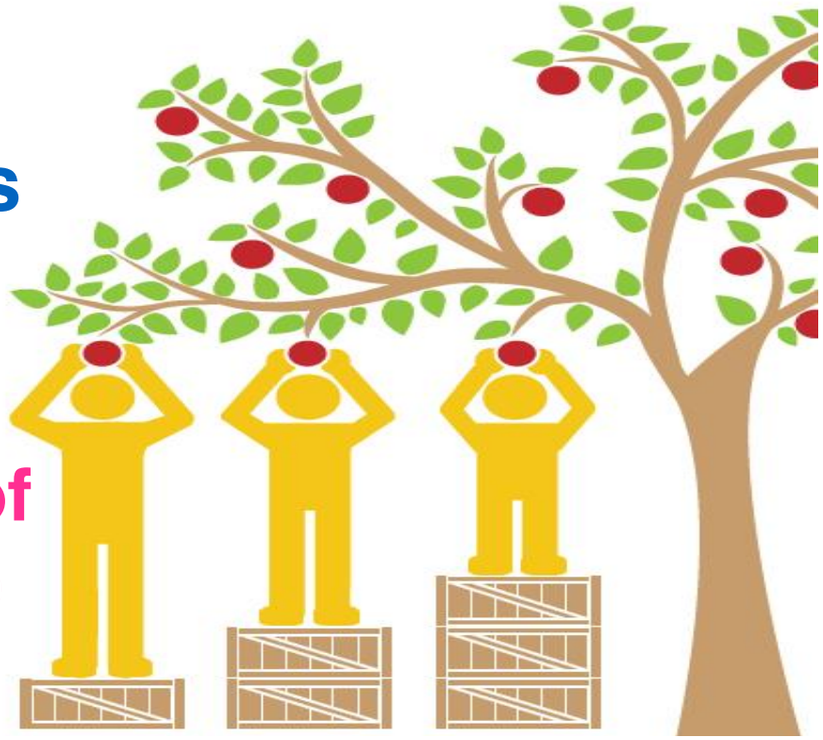
Structural bias examples: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

Fairness and worldviews

More on this in week 4

group fairness

equality of outcome



individual fairness

equality of treatment

r/ai

Fairness measures

r/ai

Fairness measures

- Fairness through unawareness
- Individual fairness
- Demographic parity
- Equalized odds
- Calibration

Review of fairness measures

Notation

A: protected attributes

X: observable attributes

U: unobserved attributes

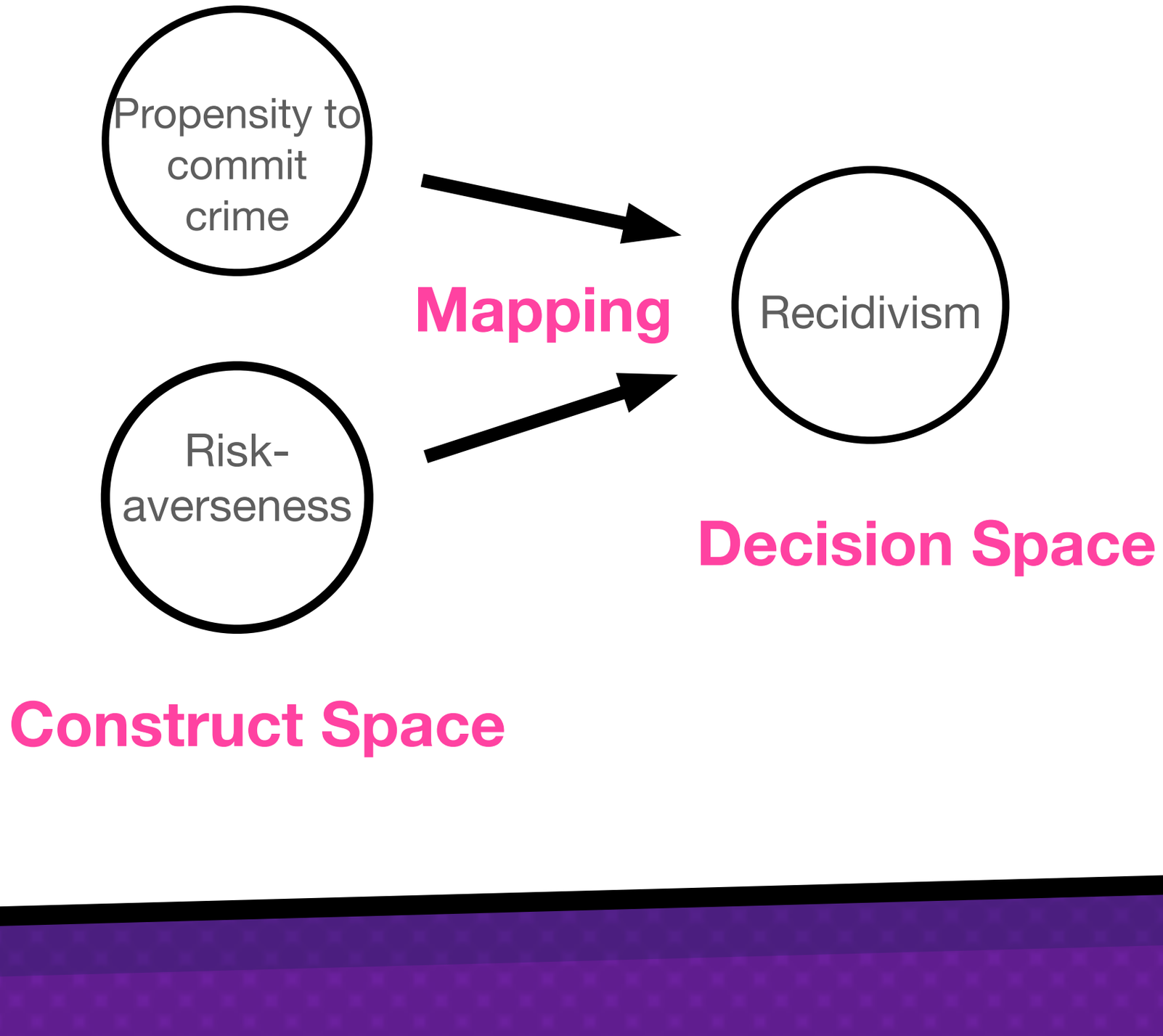
Y: outcome

$\hat{\mathbf{Y}}$: predictor (produced by a machine learning algorithm as a prediction of Y)

Capital letters refer to features and lower case letters refer to a value that feature takes

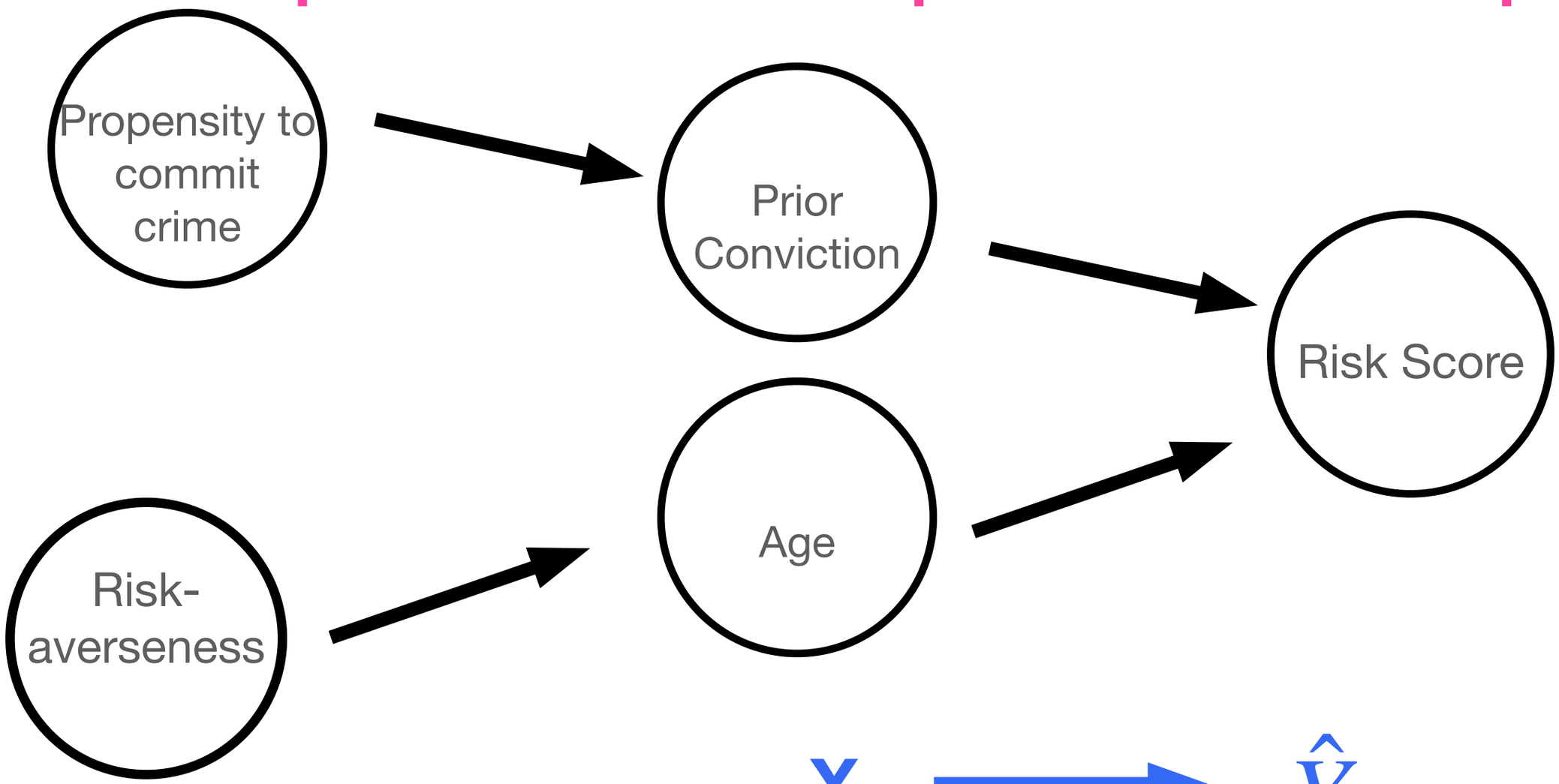
e.g. suppose **A** is age, then **a** = old and **a′** = young

Mapping CS to DS



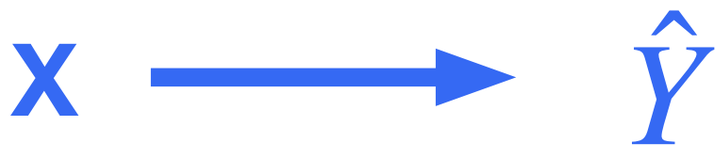
Mapping X to Y

Construct Space **Observed Space** **Decision Space**



Mapping X to Y

Many ways to map X to $\hat{Y}$



Lab 1: Mapping function was Logistic Regression

```
model = sm.formula.glm(formula = formula,
                       family = sm.families.Binomial(),
                       data = df_logistic).fit()

print(model.summary())
```

Fairness through unawareness

A predictor $\hat{Y}$ satisfies fairness through unawareness if:

$$P(\hat{Y} = y \mid X = x)$$

- Predictions do not explicitly use protected attributes, **A**

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

Seattle School

Majority opinion

"Classifying and assigning school-children according to a binary conception of race is an extreme approach."

Dissenting opinion

Today's result "undermines Brown's promise of integrated primary and secondary education that local communities have sought to make a reality."

The New York Times

The New York Times

Justices Limit the Use of Race in School Plans for Integration

Give this article

By Linda Greenhouse

June 29, 2007

WASHINGTON, June 28 — With competing blocs of justices claiming the mantle of Brown v. Board of Education, a bitterly divided Supreme Court declared Thursday that public school systems cannot seek to achieve or maintain integration through measures that take explicit account of a student's race.

Chief Justice Roberts

> "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."
>
> Chief Justice John Roberts (2007)

i.e. fairness through unawareness:

$$P(\hat{Y} = y \mid X = x)$$

- Do not explicitly use protected attributes, **A**

Individual fairness

A predictor $\hat{Y}$ satisfies individual fairness if:

$$P(\hat{Y}^i = y \mid X^i, A^i) \approx P(\hat{Y}^j = y \mid X^j, A^j)$$

$$\text{if } d(i, j) \approx 0$$

Here, **d** is a task-specific metric that measures the similarity of individuals *i* and *j*.

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Demographic parity

A predictor $\hat{Y}$ satisfies demographic parity if:

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a')$$

- Predictions are independent of A

If this is not satisfied, we have disparate impact

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Equalized odds

A predictor $\hat{Y}$ has equalized odds if:

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

- If a person truly has state **y**, the classifier will predict this at the same rate regardless of the value of **A**

$$\hat{Y} \perp\!\!\!\perp A \mid Y$$

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Equalized odds

The COMPAS predictor $\hat{Y}$ violated equalized odds. Specifically:

$$P(\hat{Y} = y \mid A = \text{Black}, Y = 0) \neq P(\hat{Y} = y \mid A = \text{White}, Y = 0)$$

- The prediction **y** for Black defendants who did not reoffend was higher than for White defendants who did not reoffend.
- Recall: FPR imbalance.

Back to ProPublica's COMPAS study

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

May 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. COMPAS has been used by the U.S. states of NY, WI, CA, FL and other jurisdictions.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

FPR

FNR

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

A more general statement: Balance

- Balance for the positive class: Positive instances are those who go on to re-offend. The average score of positive instances should be the same across groups.
- Balance for the negative class: Negative instances are those who do not go on to re-offend. The average score of negative instances should be the same across groups.
- Generalization of: Both groups should have equal false positive rates and equal false negative rates.
- Different from statistical parity!

the chance of making a mistake does not depend on race

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Calibration

A predictor $\hat{Y}$ is calibrated if:

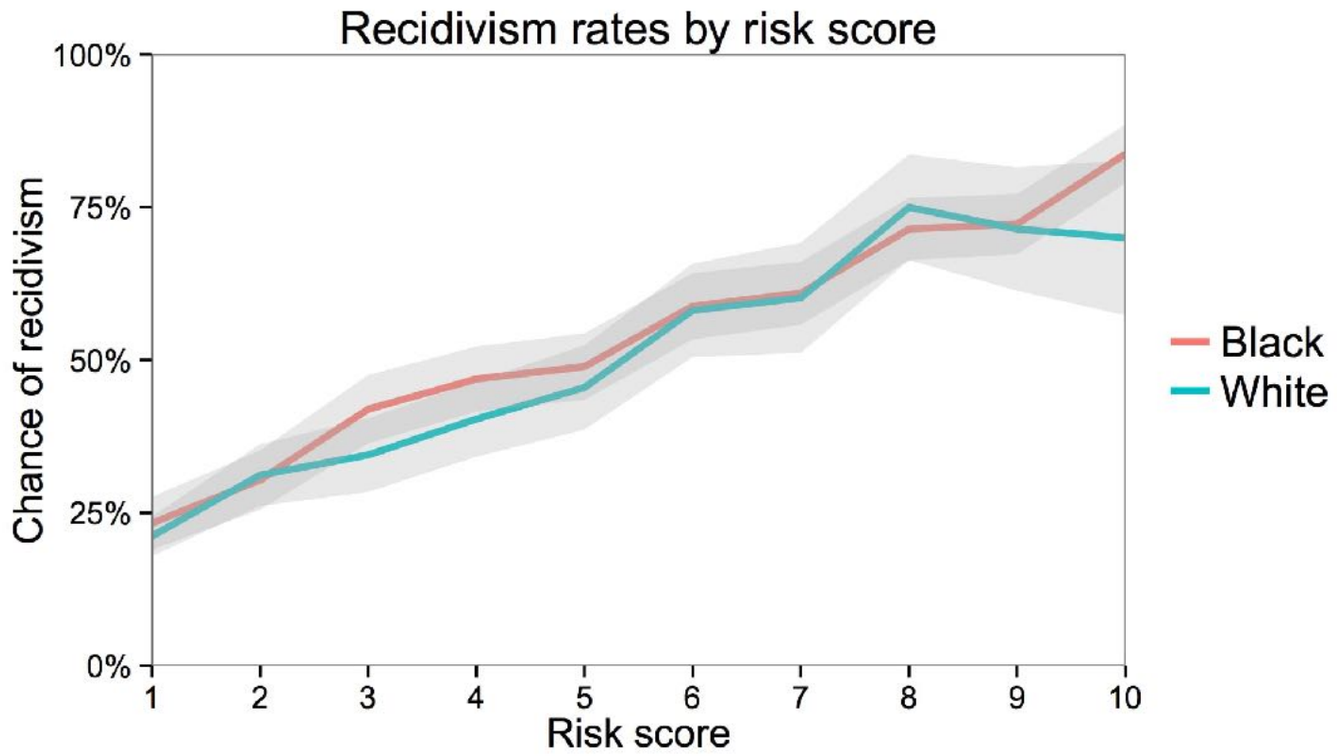
$$P(Y = y \mid A = a, \hat{Y} = y) = P(Y = y \mid A = a', \hat{Y} = y)$$

- If the classifier predicts that a person has state **y**, their probability of actually having state **y** should be the same for all values of **A**

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

COMPAS as a predictive instrument

COMPAS is reasonably well-calibrated:



[plot from Corbett-Davies et al.; *WaPo 2016*]

Calibration

The COMPAS $\hat{Y}$ is calibrated:

$$P(Y = y \mid A = \text{Black}, \hat{Y} = 0.8) = P(Y = y \mid A = \text{White}, \hat{Y} = 0.8)$$

- This sounds similar to equalized odds. But they are fundamentally incompatible
- In nearly all real cases, we cannot satisfy calibration **and** equalized odds at the same time

Calibration and Predictive Parity

Notation

S = S(x): risk score based on covariates

Calibration:

$$P(Y = 1 \mid S = s, R = b) = P(Y = 1 \mid S = s, R = w)$$

Predictive Parity:

$$P(Y = 1 \mid S > s_{HR}, R = b) = P(Y = 1 \mid S > s_{HR}, R = w)$$

[Chouldechova, *Big Data* 2017]

Racial bias in healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

October 2019

Ziad Obermeyer[1,2,*], **Brian Powers**[3], **Christine Vogeli**[4], **Sendhil Mullainathan**[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Science

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias**: **At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses**. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm **predicts health care costs rather than illness**, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, **despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise**. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

r/ai

Racial bias in healthcare

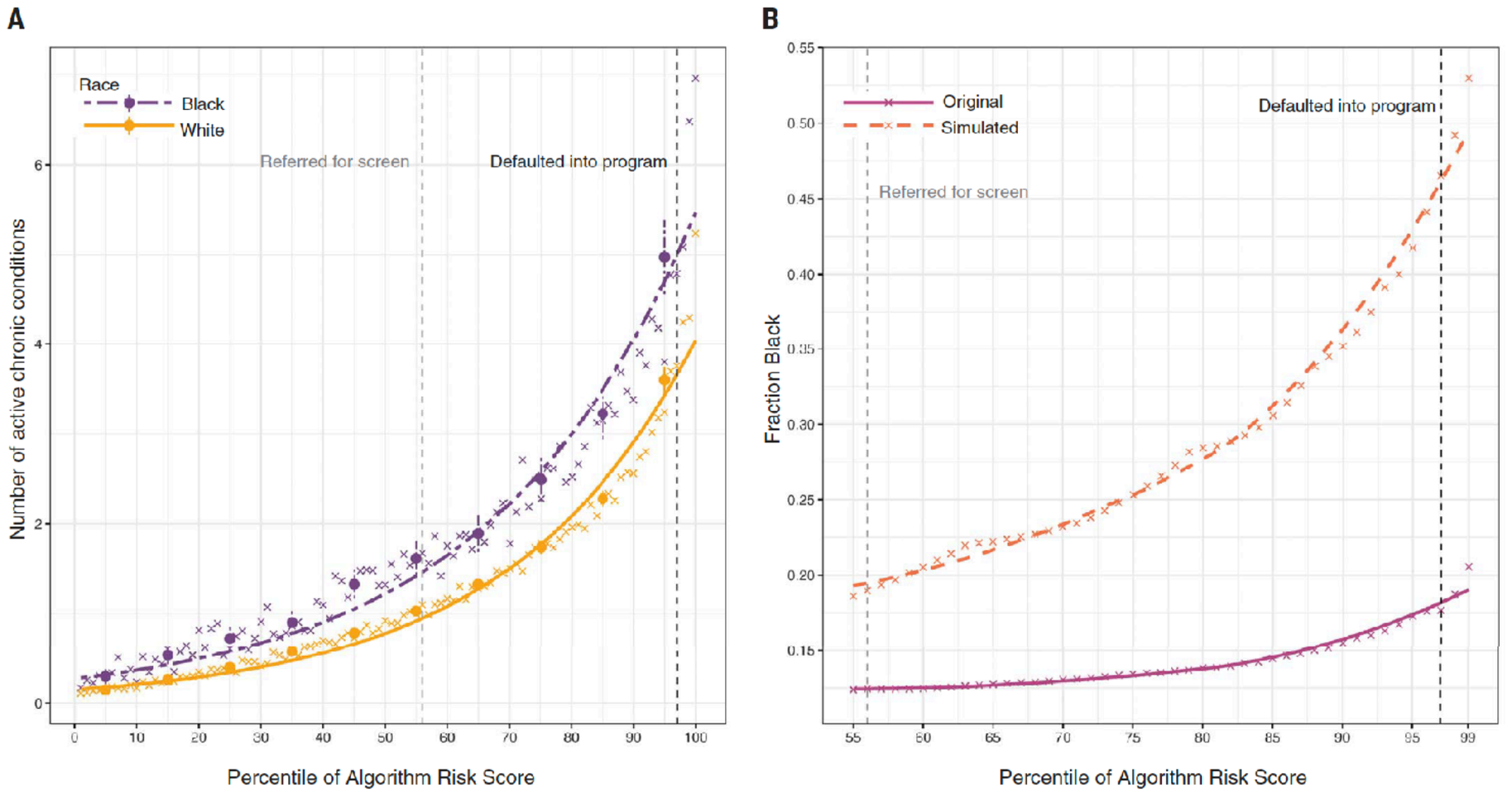


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

https://www.science.org/doi/10.1126/science.aax2342

r/ai