Responsible Data Science

Fairness and Causality

Prof. George Wood

Center for Data Science
New York University

data RESPONSIBLY

NYU

Fairness and causality

1. Review of fairness measures
2. Causal models
3. Causal models as a framework for fairness

Reading: Fairness and causality

The long road to fairer algorithms

Build models that identify and mitigate the causes of discrimination.

Matt J. Kusner & Joshua R. Loftus

Causal Reasoning for Algorithmic Fairness

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

Abstract

In this work, we argue for the importance of causal reasoning in creating fair algorithms for decision making. We give a review of existing approaches to fairness, describe work in causality necessary for the understanding of causal approaches, argue why causality is necessary for any approach that wishes to be fair, and give a detailed analysis of the many recent approaches to causality-based fairness.

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

fairness measures, a review

r/ai

Review of 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

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]

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 (2017)

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:

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

When $d(i, j) \approx 0$. 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:

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

- 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]

Demographic parity

In Lab 2, the predictor $\hat{Y}$ satisfied demographic parity after our in-processing fairness intervention:

$$P(\hat{Y} = y \mid A = \text{young}) = P(\hat{Y} = y \mid A = \text{old})$$

[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**

[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.

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]

Calibration

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

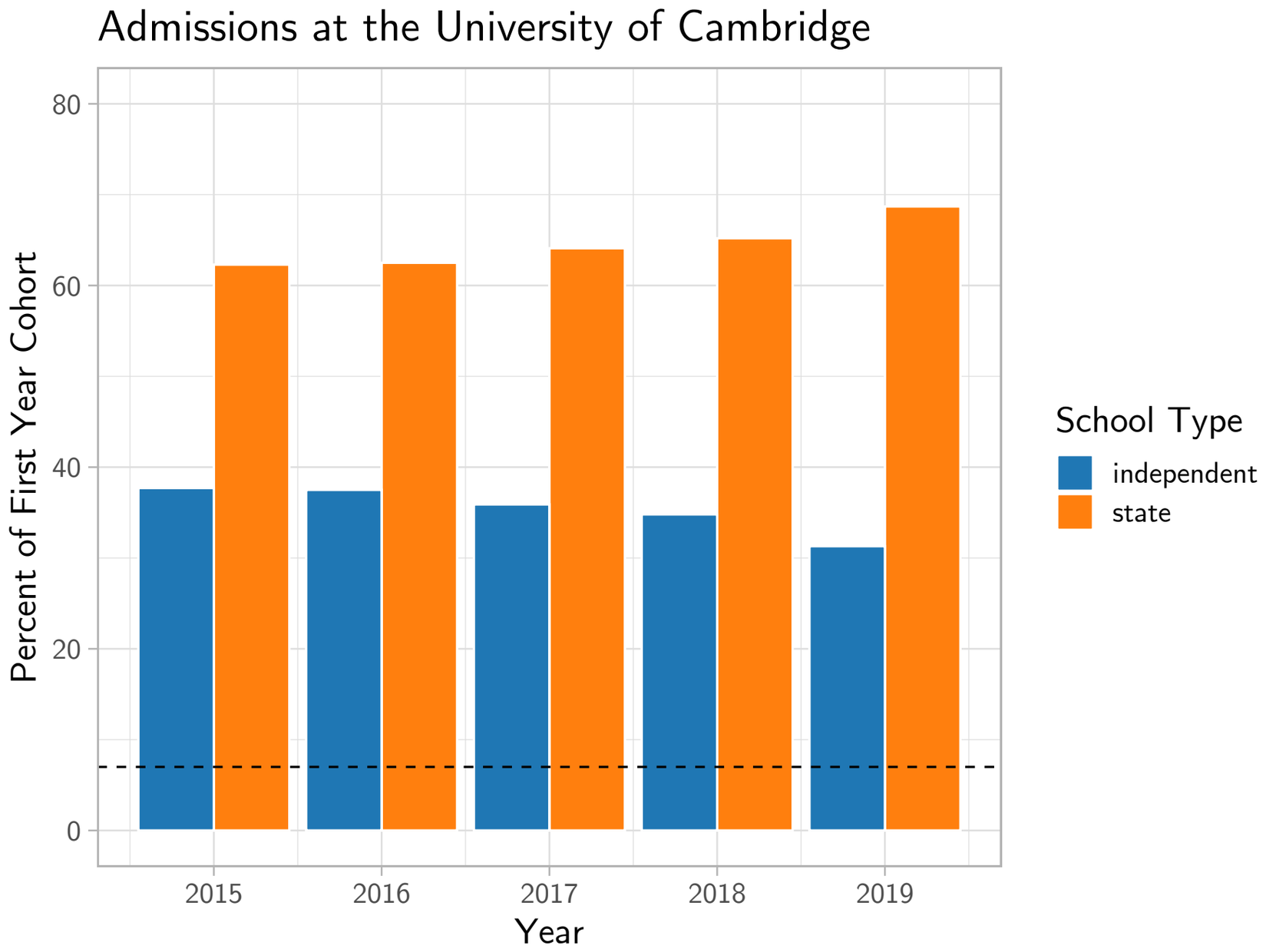
$$P(Y = y \mid A = \text{Black}, \hat{Y} = 0.8) = P(Y = y \mid A = \text{White}, \hat{Y} = 0.8)$$

- 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

causal models

r/ai

What is a causal model?



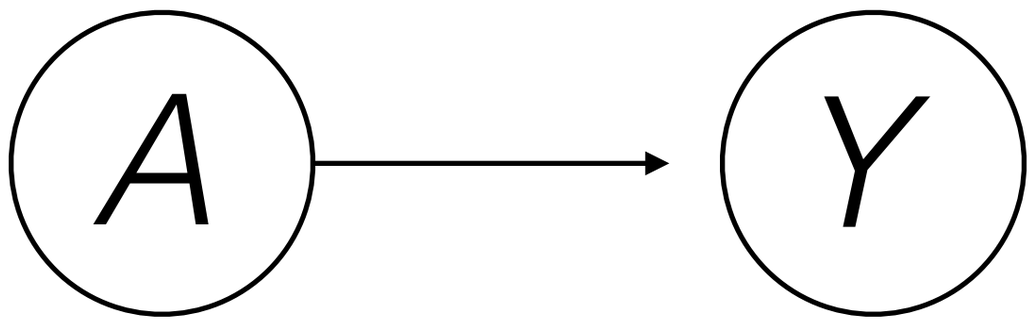
What is a causal model?

A, **Intervention**:

Student attends an independent school

Y, **Outcome**:

Student gets a place at Cambridge



We can represent this causal structure using a Directed Acyclic Graph (DAG)

What is a causal model?

A causal model presupposes a counterfactual

A, **Intervention**:

Student attends an independent school

→

Y, **Outcome**:

Student gets a place at Cambridge

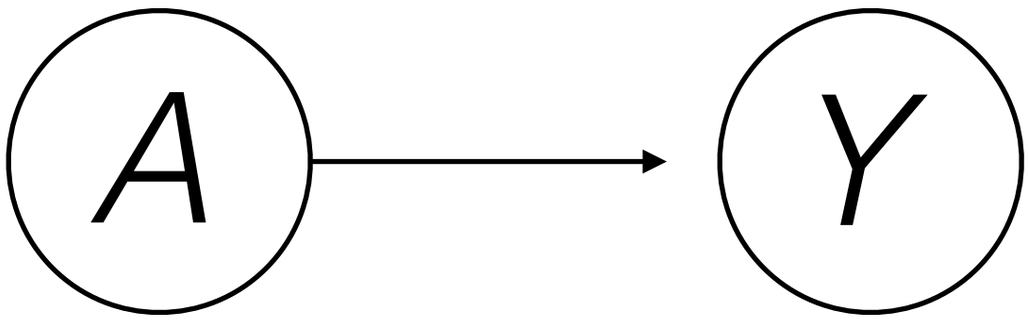
A´, **Intervention**:

Student does not attend an independent school

→

Y, **Outcome**:

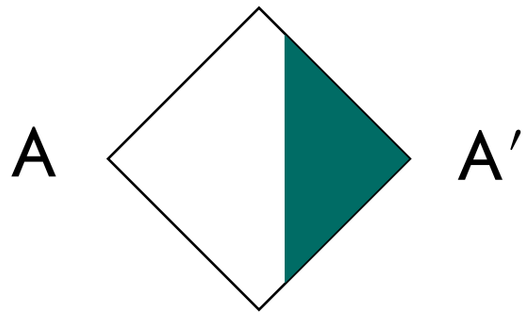
Student does not get a place at Cambridge



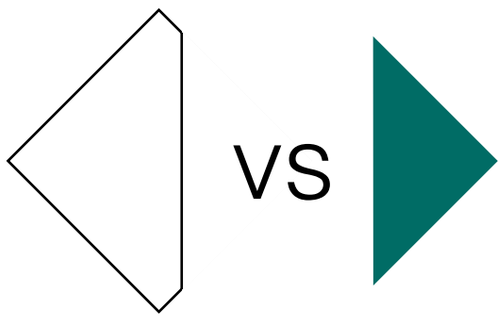
r/ai

Association and causation

Population of individuals

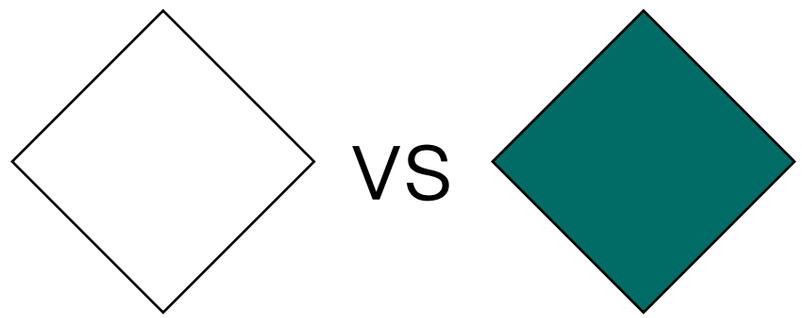


Association



$P[Y^{a=0}]$ $P[Y^{a=1}]$

Causation



$P[Y|A=0]$ $P[Y|A=1]$

r/ai

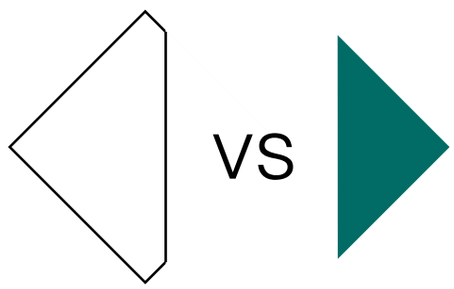
Association and causation

Population of individuals

What is the probability of going to Cambridge for students at state schools

What is the probability of going to Cambridge for students at independent schools

Association



$P[Y^{a=0}]$ $P[Y^{a=1}]$

⇒ The world as it is

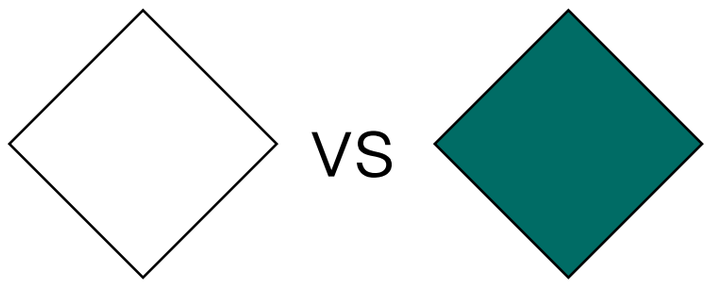
r/ai

Association and causation

Population of individuals

A′

Causation



P[$Y|A=0$] P[$Y|A=1$]

What if a student at the independent school had attended a state school?

What if a student at the state school had attended an independent school?

⇒ A counterfactual world

Fundamental problem of causal inference

We cannot observe the counterfactual

causal models and fairness

r/ai