Algorithmic bias and fairness

13 Jan 2022

an algorithm of sorts

a warm-up

- What if we think of the process of calculating a course grade as an algorithm?
- The instructor collects data (assignment grades), combines them to compute some numerical summary (course average) and finally converts them to final course grade decisions (out of 4.0).

a warm-up

- How might this “algorithm” be unfair? What are potential harms of this “algorithm?”

- Brainstorm as many potential harms as possible and potential causes of harms.

- ex. Potential harm: Students who are struggling with COVID-related difficulties may be unfairly penalized.

- ex. Potential cause: Points are given for attendance, penalizing students who test positive.

Discussion

- Is it reasonable to call this an “algorithm?”
- How might we sort/categorize potential causes of harms?
- Are all potential harms biases?

today’s plan

Questions for today

- What exactly is algorithmic bias? Algorithmic fairness?
- How do these ideas relate to statistical bias?
- How do we systematically identify and analyze ethical concerns of algorithms/algorithmic systems?

algorithmic bias

What is the ideal information the algorithm should be providing in this instance? It was supposed to identify patients who were going to get sick tomorrow so hospitals could enroll them in the extra help program today. We call that goal of the algorithm its **ideal target**. But what was the algorithm actually doing? In fact, it was doing something subtly but importantly different: it was predicting not who was going to get sick but who was going to generate high costs for the health care system. This is the algorithm's **actual target**. The wedge between these two is a key driver of bias.

Bembeneck et al.

Instead, regulators can focus on one simple question: is the algorithm predicting its ideal target accurately and equitably? This test will detect many forms of bias, like failure of an algorithm to generalize from one population to another.

Bembeneck et al.

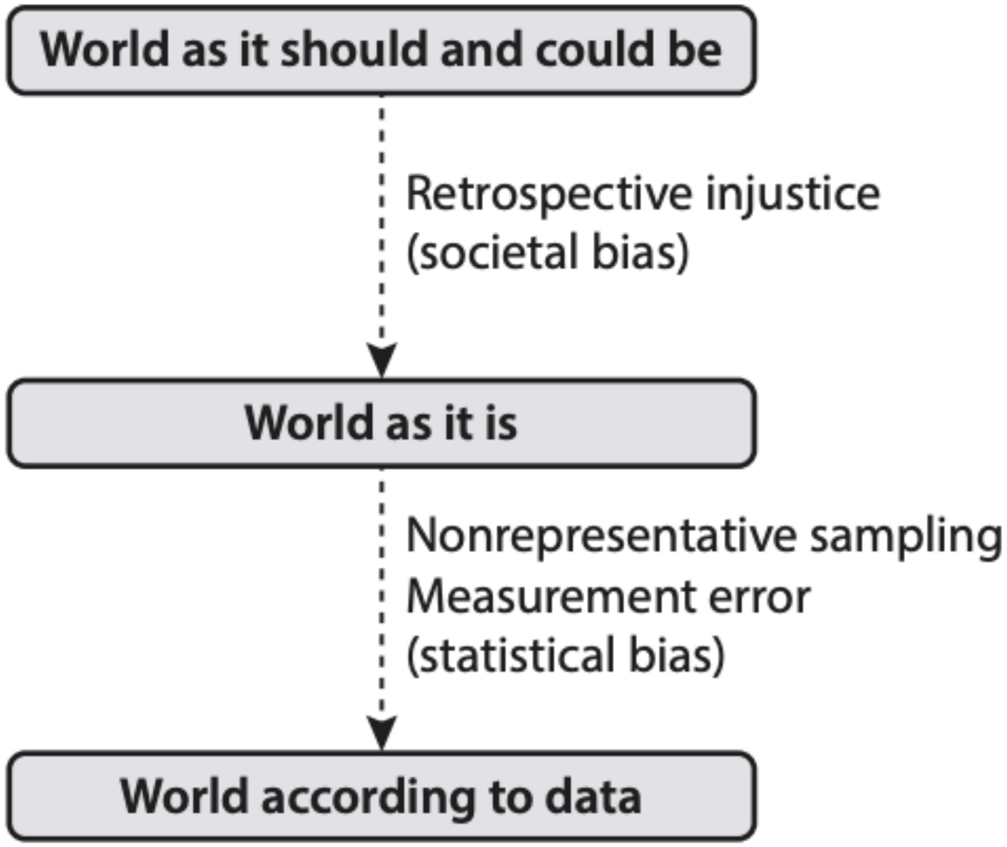
… we use the term bias to refer to computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others.

Friedman and Nissenbaum, “Bias in computer systems.”

What is algorithmic bias?

One working definition:

Algorithmic bias refers to systematic errors in an algorithm or automated decision system that lead to unfair outcomes, such as less useful results for different groups of users/subjects.



World as it should and could be

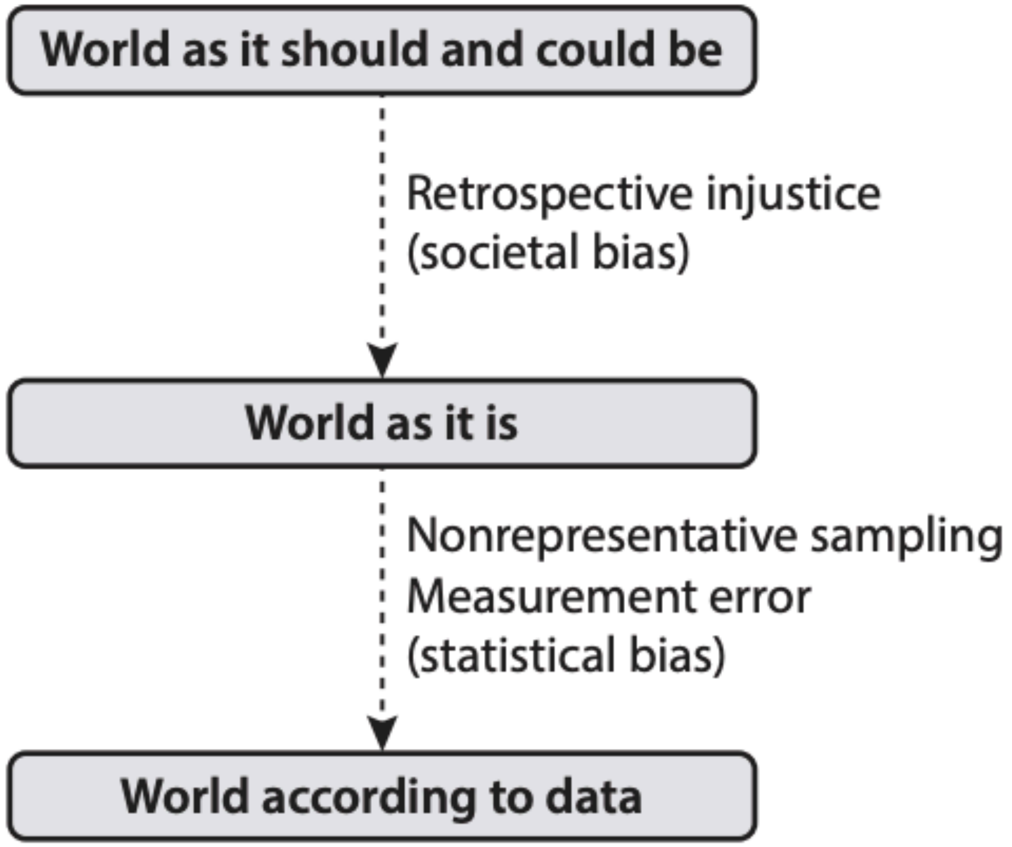
Retrospective injustice
(societal bias)

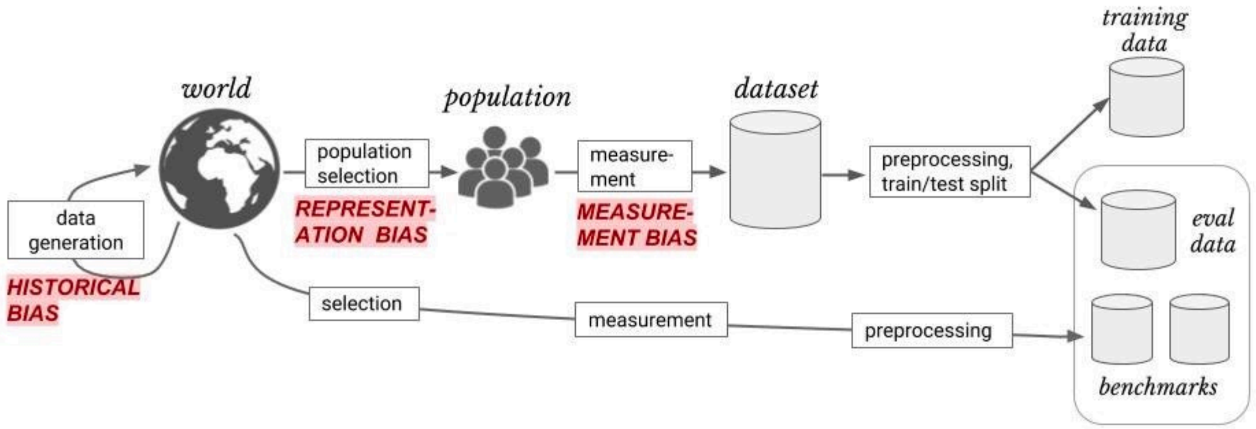
World as it is

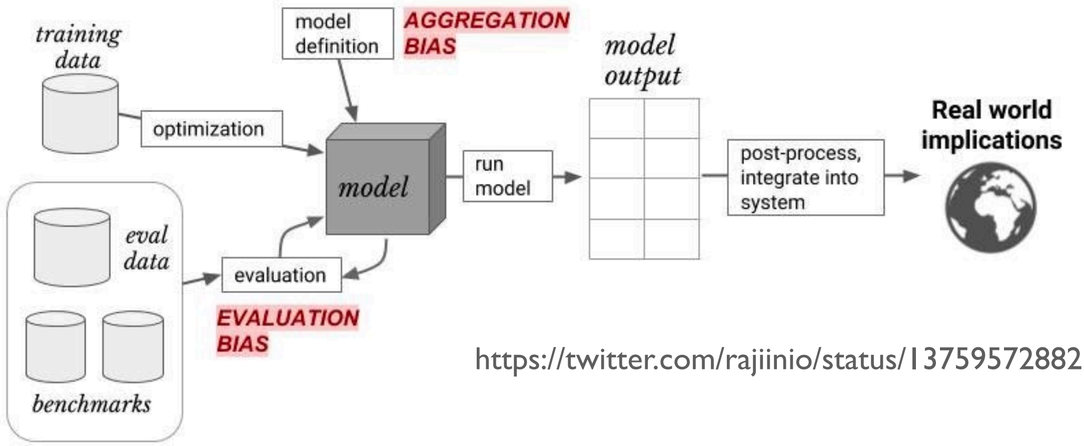
Nonrepresentative sampling
Measurement error
(statistical bias)

World according to data

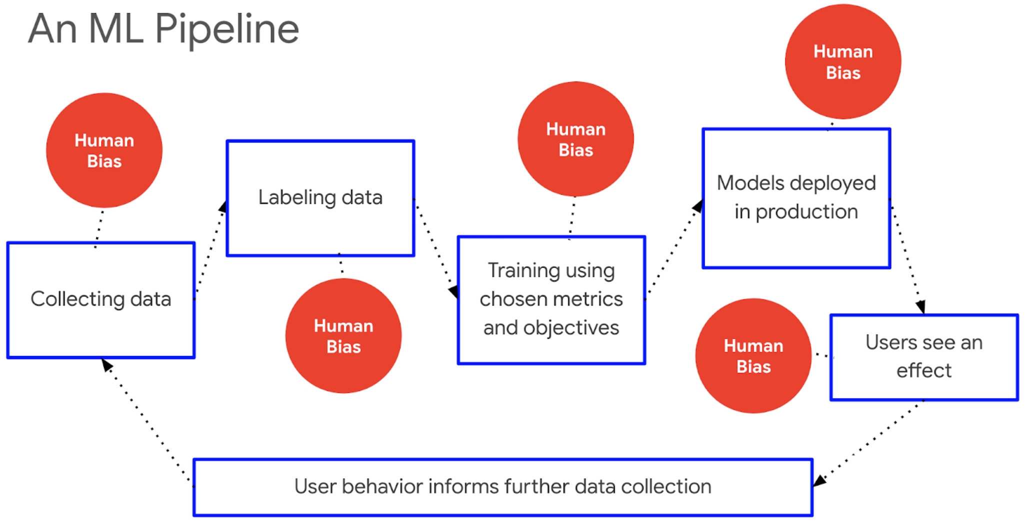
Mitchell et al.



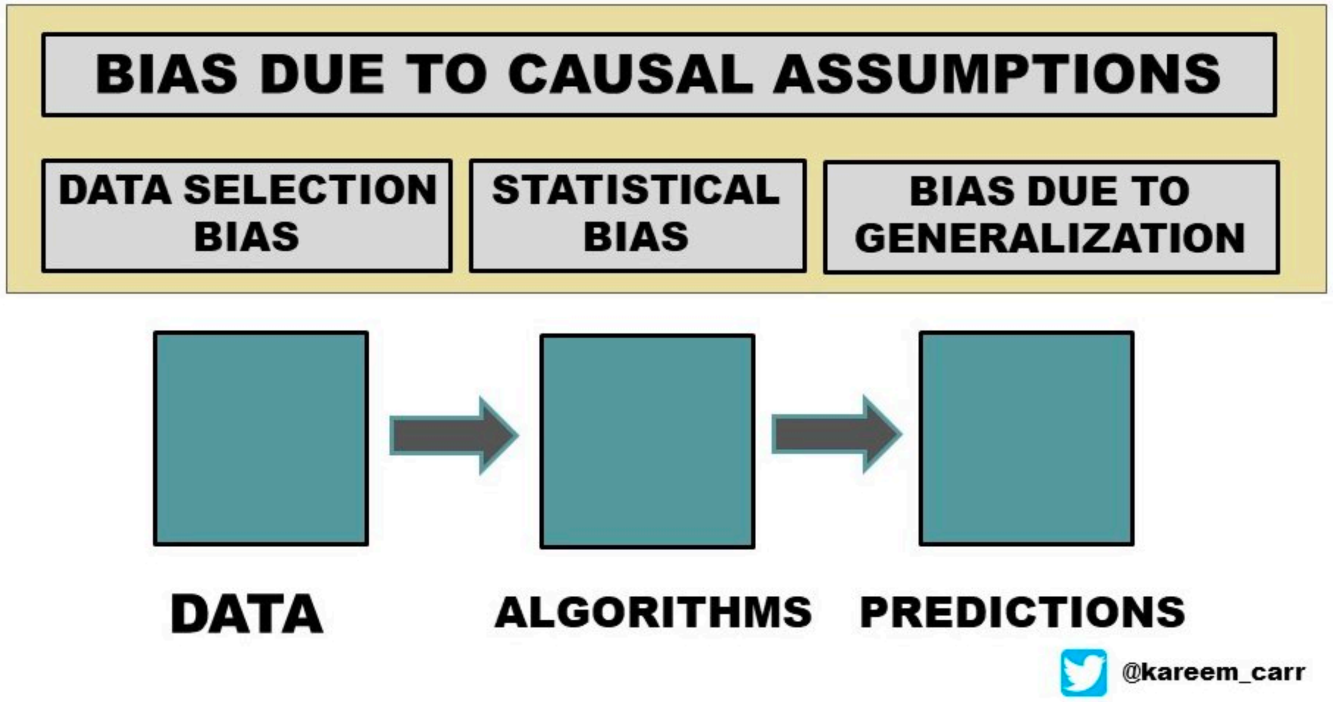




https://twitter.com/rajiinio/status/1375957288276611075


An ML Pipeline
Human Bias
Collecting data
Labeling data
Human Bias
Human Bias
Training using chosen metrics and objectives
Human Bias
Models deployed in production
Human Bias
Users see an effect
User behavior informs further data collection

https://blog.tensorflow.org/2019/12/fairness-indicators-fair-ML-systems.html



Discussion

- Is this definition of algorithmic bias adequate? Why or why not?

statistical bias

What is statistical bias?

Statistical bias refers to systematic errors in a particular estimation problem.

Usually, we use sample-based statistics to estimate population parameters.

ex. Suppose we wish to know: *What is the average height of US adults?*

We could sample 50 people and report their mean height…

What is statistical bias?

Statistical bias refers to differences between the expected results of our analysis and the true population parameter or characteristic.

In essence, statistical bias refers to systematic differences between our estimate and what we want to estimate.

The question: What is it exactly that we want to estimate? is tricky though.

What is statistical bias?

Imagine we wish to design an algorithm that estimates the probability that a college basketball player will succeed in the NBA.

On average, the algorithm gives Duke players a higher probability of success than WSU players. Is this statistical bias?

Discussion

- What are potential sources of statistical bias?
- What is the relationship between statistical bias and algorithmic bias?

algorithmic fairness

> “In recent years, attention has focused on how consequential predictive models may be biased—a now overloaded word that, in popular media, has come to mean that the model's performance (however defined) unjustifiably differs along social axes such as race, gender, and class. Uncovering and rectifying such biases in statistical and machine learning models has motivated a field of research we call algorithmic fairness.”

Mitchell et al. “Algorithmic Fairness”

What about algorithmic justice?

Are fair algorithms just?

One potential hole in the terms "bias" and "fairness" comes from the idea that what is fair from one perspective is not fair from another perspective.

Discussion

- How should we determine what is fair? How do we determine what is fair?
- Is ensuring algorithmic fairness enough?

a systematic study of algorithmic systems?

ShotSpotter

The company's algorithms initially classified the sound as a firework. That weekend had seen widespread protests in Chicago in response to George Floyd's murder, and some of those protesting lit fireworks.

But after the 11:46 p.m. alert came in, a ShotSpotter analyst manually overrode the algorithms and "reclassified" the sound as a gunshot. Then, months later and after "post-processing," another ShotSpotter analyst changed the alert's coordinates to a location on South Stony Island Drive near where Williams' car was seen on camera.

Todd Feathers, "Police Are Telling ShotSpotter to Alter Evidence From Gunshot-Detecting AI"

The company has not allowed any independent testing of its algorithms, and there's evidence that the claims it makes in marketing materials about accuracy may not be entirely scientific.

Over the years, ShotSpotter's claims about its accuracy have increased, from 80 percent accurate to 90 percent accurate to 97 percent accurate. According to Greene, those numbers aren't actually calculated by engineers, though.

Todd Feathers, "Police Are Telling ShotSpotter to Alter Evidence From Gunshot-Detecting AI"

Discussion

- What are potential harms of the Shotspotter algorithm? Are these potential biases?
- Is this algorithm biased? In what sense?
- Can this algorithm be made fair?

TikTok

The document says watch time isn’t the only factor TikTok considers. The document offers a rough equation for how videos are scored, in which a prediction driven by machine learning and actual user behavior are summed up for each of three bits of data: likes, comments and playtime, as well as an indication that the video has been played:

Plike X Vlike + Pcomment X Vcomment + Eplaytime X Vplaytime + Pplay X Vplay

“The recommender system gives scores to all the videos based on this equation, and returns to users videos with the highest scores,” the document says. “For brevity, the equation shown in this doc is highly simplified. The actual equation in use is much more complicated, but the logic behind is the same.”

Ben Smith, “How TikTok Reads Your Mind”

Julian McAuley, a professor of computer science at the University of California San Diego, who also reviewed the document, said in an email that the paper was short on detail about how exactly TikTok does its predictions, but that the description of its recommendation engine is "totally reasonable, but traditional stuff." The company's edge, he said, comes from combining machine learning with "fantastic volumes of data, highly engaged users, and a setting where users are amenable to consuming algorithmically recommended content (think how few other settings have all of these characteristics!). Not some algorithmic magic."

Ben Smith, "How TikTok Reads Your Mind"

Discussion

- What are potential harms of the "TikTok algorithm?" Are these potential biases?
- Is the "TikTok algorithm" really reading your mind?
- Is this algorithm biased? In what sense?
- Can this algorithm be made fair?

Discussion

- How can we relate these terms:
  - algorithmic bias
  - algorithmic fairness
  - statistical bias
  - algorithmic harms

Discussion

- What are some of the questions we need to ask about algorithms and automated decision systems?
- How do we systematically identify problems?