Bias in Algorithms

Professor Joel Shapiro
Managerial Economics and Decision Sciences

Northwestern | Kellogg

PREDICTIVE MODELS ALLOW FOR DIFFERENTIATION

You are in charge of operations at a wind farm.

Your predictive model tells you there that the highest risk of turbine failure occurs in these conditions:

- Carbon fiber, (not Kevlar) and
- Horizontal axis (not vertical) and
- Manufactured in Chicago plant and
- Average humidity > 34%

Let's avoid using the turbines that are likely to fail!

PREDICTIVE MODELS ALLOW FOR DIFFERENTIATION

You are in charge of operations at a wind farm.

Your predictive model tells you there that the highest risk of turbine failure occurs in these conditions:

- Carbon fiber, (not Kevlar) and
- Horizontal axis (not vertical) and
- Manufactured in Chicago plant and
- Average humidity > 34%

Let's avoid using the turbines that are likely to fail!

You oversee heart transplants at a major medical center.

Your predictive model tells you there that the highest risk of heart failure occurs in these conditions:

- Race x
- Income level y
- County z
- Works > 1 job

Let's avoid giving the transplant to someone who is unlikely to be successful!

PREDICTIVE MODELS ALLOW FOR DIFFERENTIATION

Is this only difference that one of these is machines and one is people?

Your predictive model tells you there that the highest risk of turbine failure occurs in these conditions:

- Carbon fiber, (not Kevlar) and
- Horizontal axis (not vertical) and
- Manufactured in Chicago plant and
- Average humidity > 34%

Let's avoid using the turbines that are likely to fail!

Your predictive model tells you there that the highest risk of heart failure occurs in these conditions:

- Race x
- Income level y
- County z
- Works > 1 job

Let's avoid giving the transplant to someone who is unlikely to be successful!

DIFFERENTIATION IS NOT NECESSARILY BIAS

Businesses differentiate customers all the time. In fact, many invest in analytics to achieve their GOAL of differentiation.

- Harrah's offered only certain customers a free steak dinner
- You see different ads on Amazon than I do
- You may see different prices on Amazon than I do
- Netflix shows me different suggestions than it shows my wife

Sometimes, this differentiation starts to "feel wrong."

- I am a loyal American Airlines customer. I travel about 1x/month and almost always fly AA.
- Lilah has been a loyal AA customer, but isn't so much any more. Her travel on AA starts to decrease, and so AA flags her as "at risk," and starts offering her discounts and promotions to fly with them.

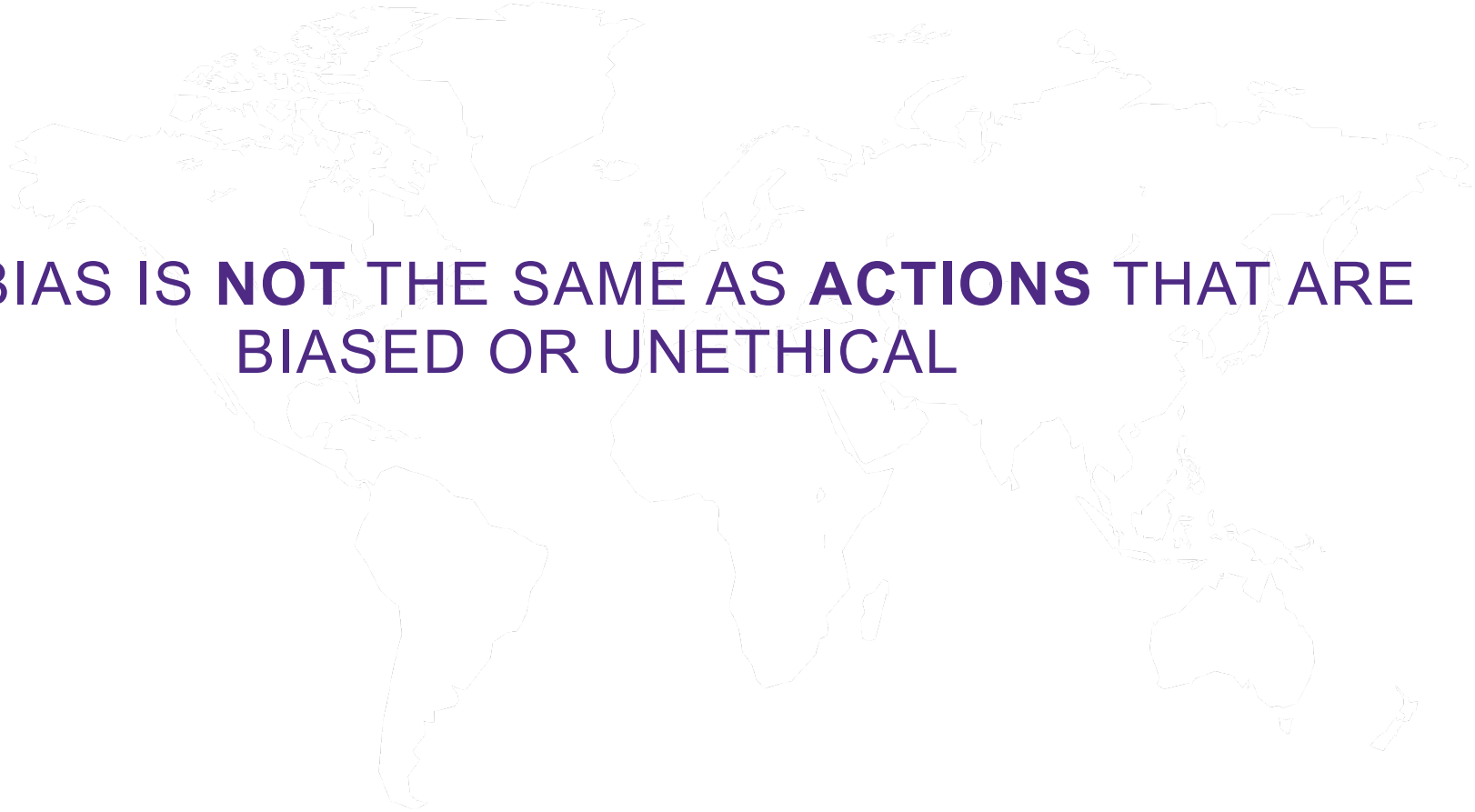
DIFFERENTIATION IS NOT NECESSARILY BIAS

DO YOU FEEL GOOD ABOUT OFFERING BETTER PRICES TO "WORSE" CUSTOMERS?

This might be unethical. Or it might be bad business. But the model is not biased.

Sometimes, this differentiation starts to "feel wrong."

- I am a loyal American Airlines customer. I travel about 1x/month and almost always fly AA.
- Lilah has been a loyal AA customer, but isn't so much any more. Her travel on AA starts to decrease, and so AA flags her as "at risk," and starts offering her discounts and promotions to fly with them.



ALGO BIAS IS **NOT** THE SAME AS **ACTIONS** THAT ARE BIASED OR UNETHICAL

WHAT **IS** ALGO BIAS?

WORLD ECONOMIC FORUM

Artificial intelligence and bias

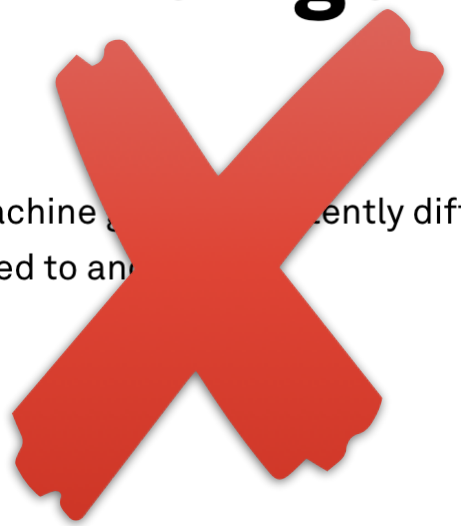
Bias in AI is when the machine gives consistently different outputs for one group of people compared to another.

WHAT **IS** ALGO BIAS?

WORLD ECONOMIC FORUM

Artificial intelligence and bias

Bias in AI is when the machine [illegible]ently different outputs for one group of people compared to an[illegible]



WHAT **IS** ALGO BIAS?

WORLD ECONOMIC FORUM

Artificial intelligence and bias

> **Bias in AI algorithms can emanate from unrepresentative or incomplete training data or the reliance on flawed information that reflects historical inequalities. If left unchecked, biased algorithms can lead to decisions which can have a collective, disparate impact on certain groups of people even without the programmer's intention to discriminate.**

4 KINDS OF ALGO BIAS

Differences in Precision

I train a model to recognize faces by giving it 9,500 images of men and 500 images of women.

Result: algorithm will be more accurate for men than for women

Perpetuation of Individual Prejudices

A sexist manager disciplines and fires more men/women because s/he holds prejudices against that group. Data about employee productivity and success fed into machine.

Result: algorithm will show that group as less productive / successful

4 KINDS OF ALGO BIAS

Data Collection Asymmetries

A police department suspects that a particular area of the city is more likely to experience crime, so they allocate more enforcement resources, which finds more crime. Not because of greater crime rate, but just greater attention paid.

Result: algorithm will show greater crime in select areas

Systemic and Societal Inequities

A credit card company feeds historical lending data into a machine to predict who is most likely to pay back the loan. Loan frequency and amount is an important predictor, but a group(s) have historically had less opportunity to take and repay loans.

Result: algorithm will say that group is less credit-worthy

4 KINDS OF ALGO BIAS

"Any algorithm, that intentionally or not results in discriminatory treatment of women or any other protected class of people violates New York law."

- NY Dept of Financial Services

"We look at an individual's income and an individual's creditworthiness, which includes facts like personal credit scores, how much debt you have, and how that debt has been managed.

In all cases, we have not and will not make decisions based on factors like gender."

- Goldman Sachs

YOUR EXAMPLES

1. Why is the example you've identified bad? (Not every case of differentiation = discrimination!)
2. What do you think are the sources of biased / discriminatory outputs in your example?
3. What might be reasonable and effective bias mitigation strategies?

KEY CONSIDERATIONS ON ALGO BIAS

- Bias in an algorithm is not the same as biased / unethical actions.
- You can't recognize bias simply from model results. Just because a subgroup is identified as differently valuable, productive, healthy, etc. does not *necessarily* indicate bias.
- You can only find algo bias by examining the underlying data and how those data are fed into a model.
- Because there is still a gap between content / context experts and model-builders, we are still learning how biases enter our algorithms.
- You are not obligated to adhere to the results of a model if it provides an output with which you are uncomfortable. You may find the results of a model *inequitable* even if there is no obvious algo bias.

BEST BOOK I KNOW ON THIS TOPIC…