Algorithmic Biases and Fairness

Swati Mishra

Applications of Machine Learning (4AL3)

Fall 2024

McMaster University
ENGINEERING

Review

- Revisited Bias Variance Trade Off
- Model Building Pipeline
- Evaluation: Subset Selection, Forward and Backward step-wise Selection
- Regularization (L1 and L2)

Bias

- **Bias** is a disproportionate weight *in favor of* or *against* an idea or thing, usually in a way that is inaccurate, or unfair.

Source: Wikipedia

Bias

- **Bias** is a disproportionate weight *in favor of* or *against* an idea or thing, usually in a way that is inaccurate, or unfair.

- In science and engineering, a bias is a systematic error and is the difference between a measured value of a quantity and its unknown true value.

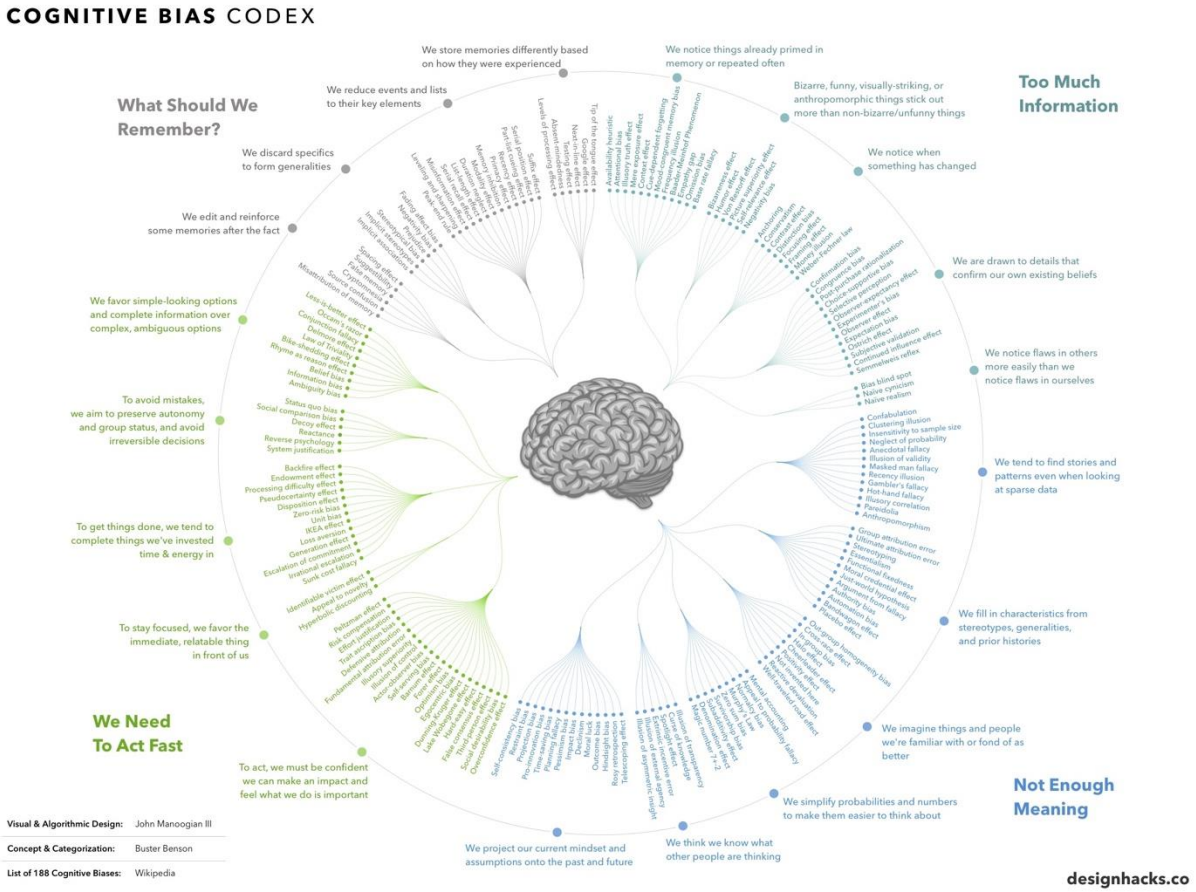
Source: Wikipedia

Bias

- **Bias** is a disproportionate weight *in favor of* or *against* an idea or thing, usually in a way that is inaccurate, or unfair.
- In science and engineering, a bias is a systematic error and is the difference between a measured value of a quantity and its unknown true value.
- Statistical bias, in the mathematical field of statistics, is a systematic tendency in which the methods used to gather data and generate statistics present an inaccurate, skewed or biased depiction of reality.

Source: Wikipedia

Types of Biases

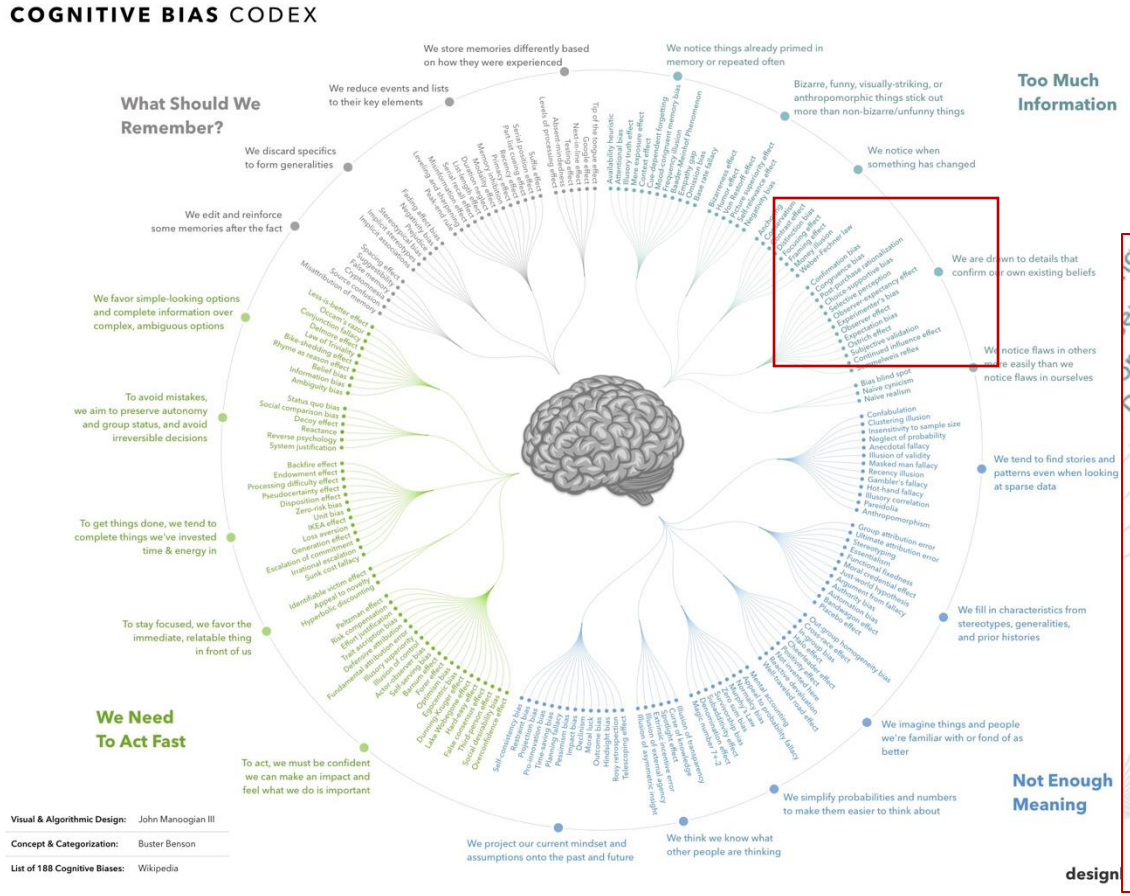


- A cognitive bias is a repeating or basic misstep in thinking, assessing, recollecting, or other cognitive processes

Source: https://www.visualcapitalist.com/every-single-cognitive-bias/

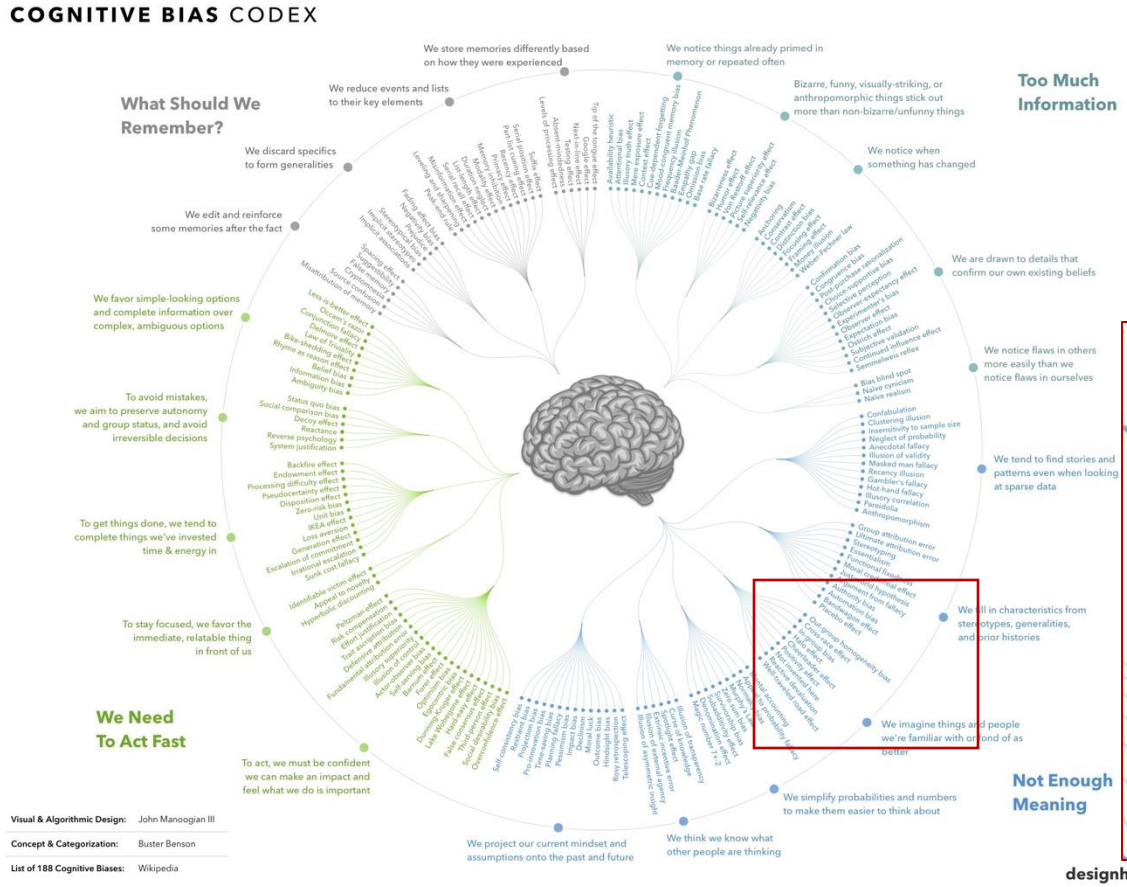
McMaster University

Types of Biases

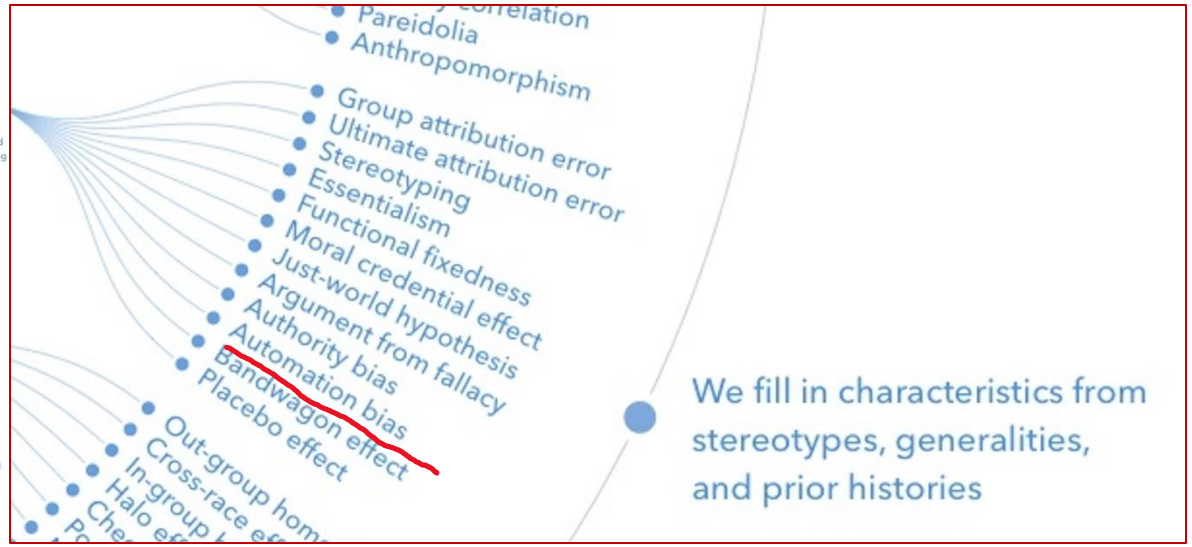


Confirmation Bias: Confirmation bias is the tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypothesis while giving disproportionately less attention to information that contradicts it.

Types of Biases

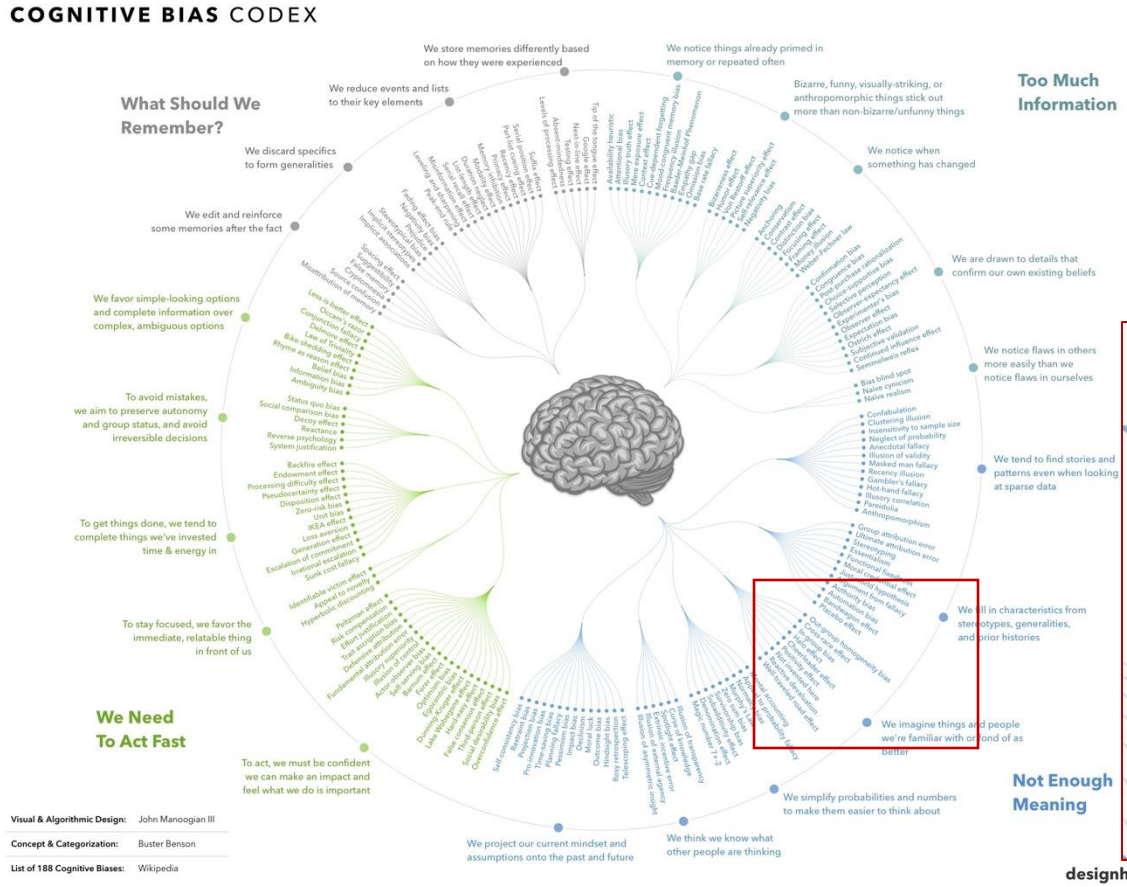


Automation bias: It is the human tendency to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.

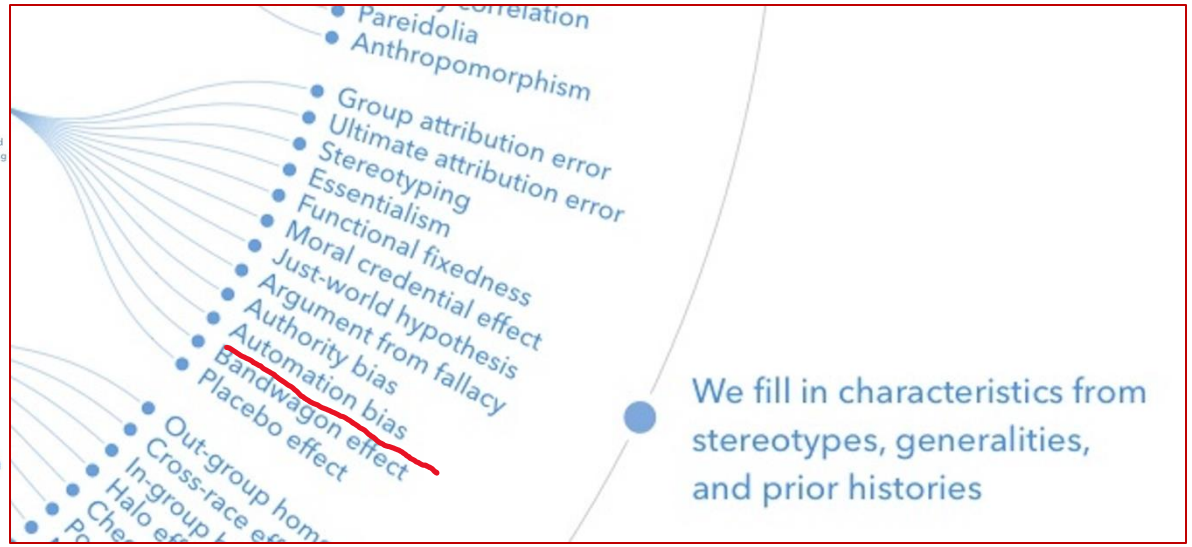


designhacks.co

Types of Biases



Automation bias: It is the human tendency to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.



designhacks.co

Types of Biases

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Gender bias in teaching evaluations: the causal role of department gender composition

Oriana R. Aragón [a,b,1], Evava S. Pietri [c], and Brian A. Powell [d,e]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received October 7, 2021; accepted December 8, 2022

January 17, 2023 | 120 (4) e2118466120 | https://doi.org/10.1073/pnas.2118466120

Significance

Women's underrepresentation at higher levels in academia negatively impacts the course of scholarly activities. One criterion for judging who reaches higher levels of academia is student-provided teaching evaluations. Here, we illustrate that both men and women suffered from gender-based discrimination in their teaching evaluations. However, women were more frequently impacted because of their minority status. When students observed gender disparities in academic departments in our experimental paradigm, they formed expectations about gendered roles for upper- and lower-level courses. Violating these expectations caused gender-based discrimination in teaching evaluations. Left unaddressed, these gender biases can undermine diversity, equity, and inclusion efforts. This article provides evidence to support those who seek to implement interventions that lead to equity and gender parity.

Women are underrepresented in academia's higher ranks. Promotion oftentimes requires positive student-provided course evaluations. At a U.S. university, both an archival and an experimental investigation uncovered gender discrimination that affected both men and women. A department's gender composition and the course

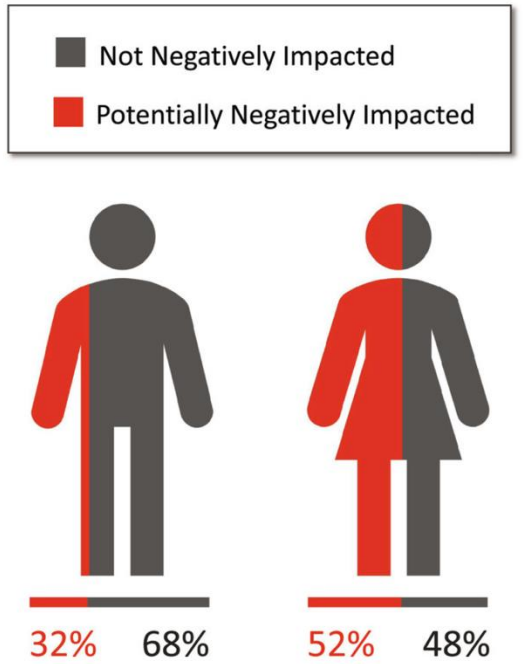


Fig. 2. In study 1, (N = 4,700), because the majority of women held gender-minority status (72.6% of departments were male dominated), and most classes taught were upper-level courses (72.3%) the potentially negative impact of gender bias was greater for women than for men. An estimated 32% of men and 52% of women at this university were potentially negatively impacted by gender bias in their teaching evaluations.

McMaster University

Types of Biases

Table 1.

Summary statistics of sections.

Course	Number of sections	Number of instructors	Female instructors (%)
Overall	1,194	379	33.8
History	230	72	30.6
Political Institutions	229	65	20.0
Microeconomics	230	96	38.5
Macroeconomics	230	93	34.4
Political science	137	49	32.7
Sociology	138	56	46.4

Data for a section of political institutions that had an experimental online format are omitted. Political science and sociology originally were not in the triad system; students were randomly assigned by the administration to different sections.

US randomized experiment

Student evaluations of teaching (mostly) do not measure teaching effectiveness

PUBLISHED ORIGINAL ARTICLE

Author(s): Anne Boring [1,2], Kellie Ottoboni [3], Philip B. Stark [*,3]

Publication date Control: 07 January 2016
Journal: ScienceOpen Research
Publisher: ScienceOpen
Keywords: Assessment, Evaluation & Research methods, Labor law, Nonparametric Statistics, Disparate Impact, Gender Bias, Permutation Tests

Print Download Review Bookmark Share Cite as...

THE FRENCH NATURAL EXPERIMENT

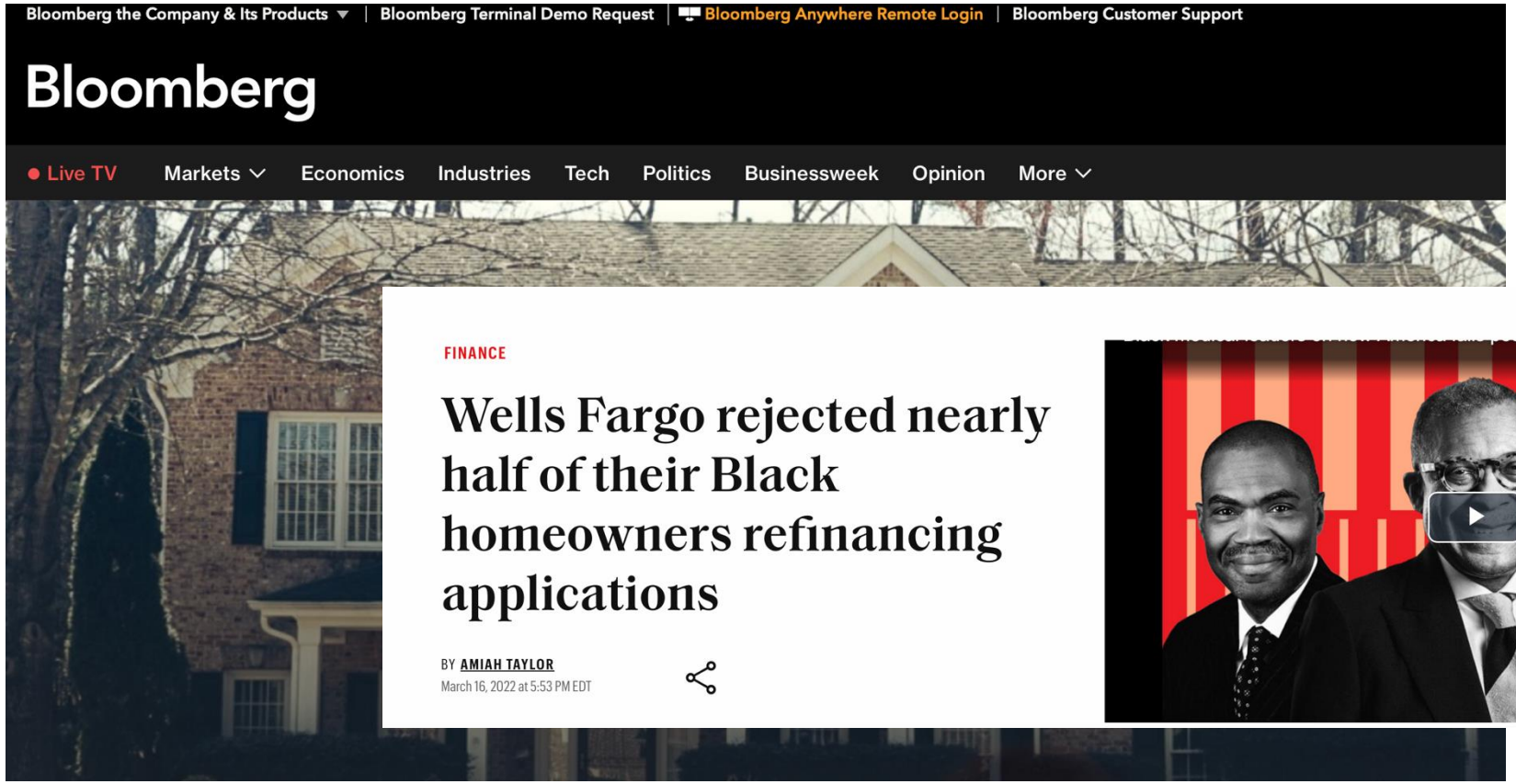
In this section, we test hypotheses about relationships among SET, teaching effectiveness, grade expectations, and student and instructor gender. Our tests aggregate data within course sections, to match how SET are typically used in personnel decisions. We use the average of Pearson correlations across strata as the test statistic,[6] which allows us to test both for differences in means (which can be written as correlations with a dummy variable) and for association with ordinal or quantitative variables.

McMaster University

Types of Biases

So what everyone is biased! Stop complaining! Why does it matter?

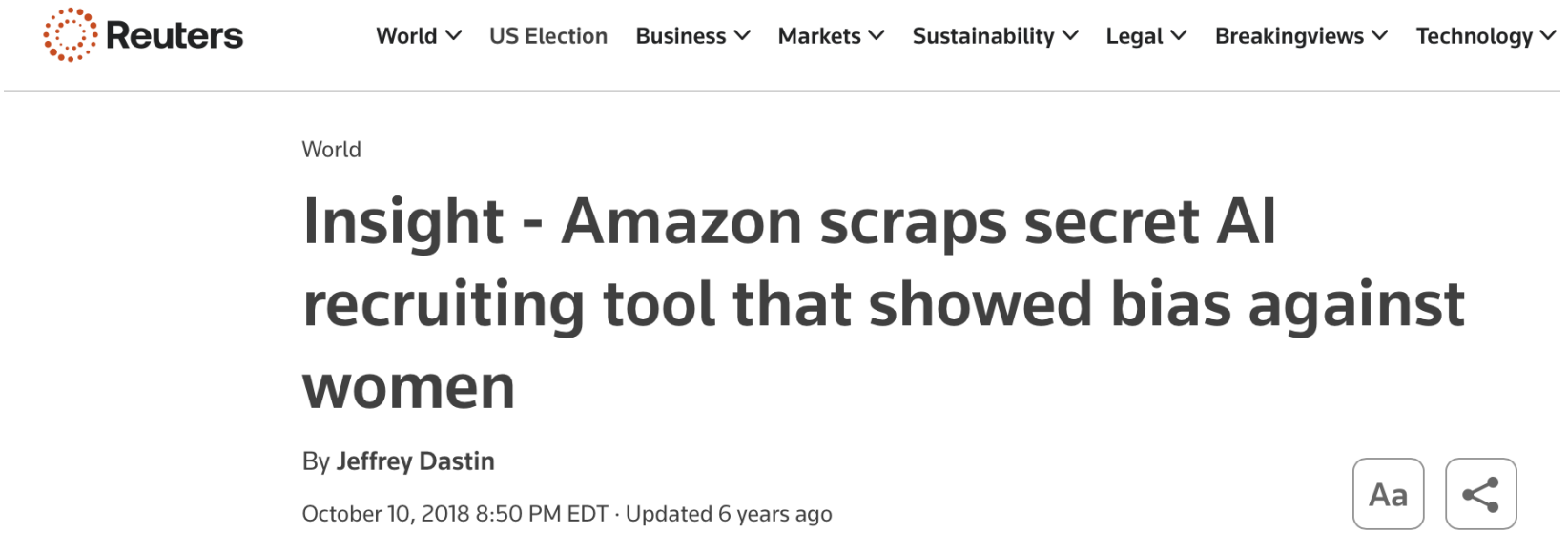
Algorithmic Biases



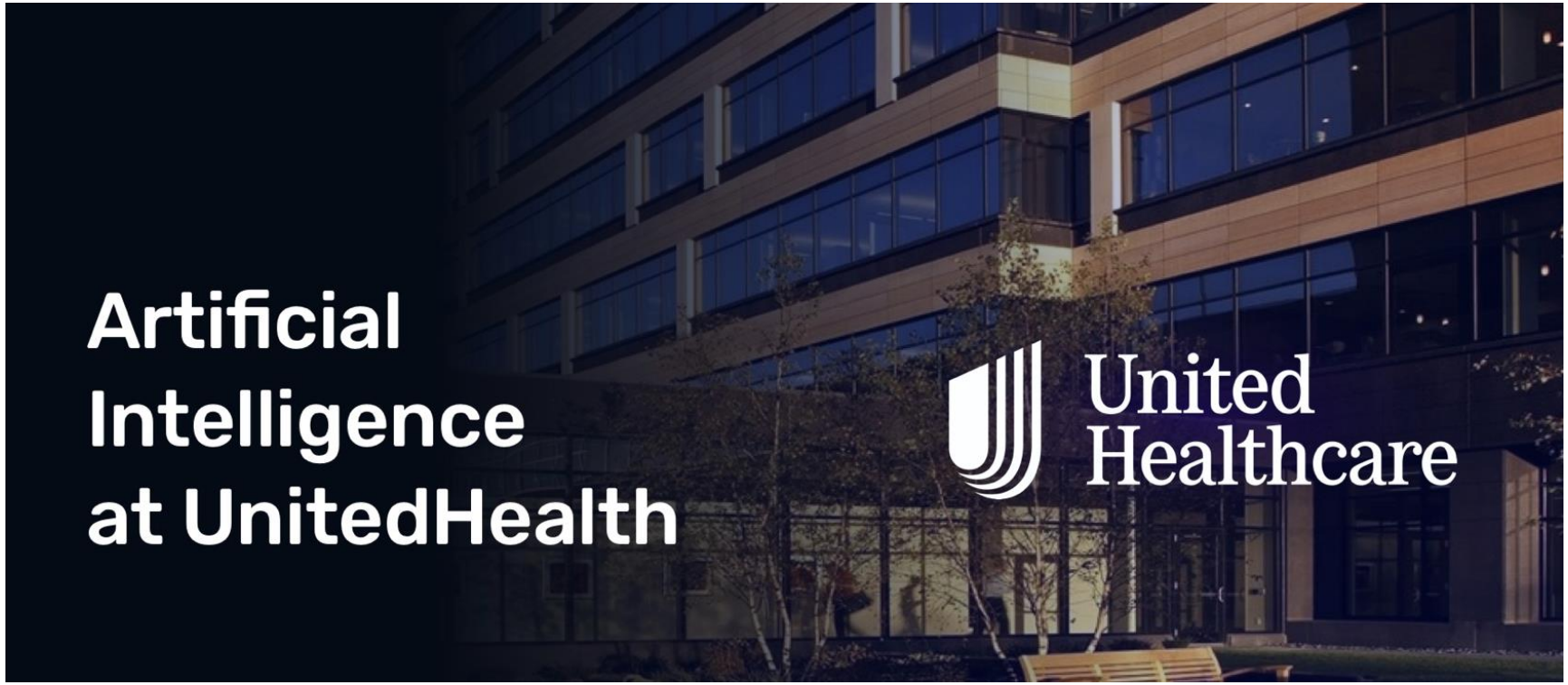
Flitter 2022, Donnan et al. 2022

McMaster University

Algorithmic Biases



Algorithmic Biases



Algorithmic Biases

- Human bias is applied far less systematically than machine bias.
- Bias can cascade with machine learning.
- Human decisions, even when biased, are often tempered within reasonable bounds, unlike machines, which may spiral out of control.
- Explaining bias in black box algorithms may appear to be hard but is feasible. Asking a human to explain how they arrived at a decision is often much harder.

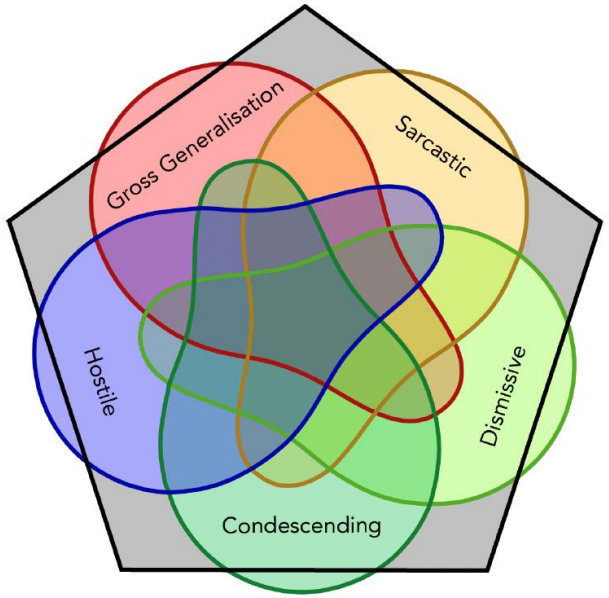
Source: Algorithmic Fairness Sanjiv Das, Richard Stanton, and Nancy Wallace

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.

What causes algorithmic biases?

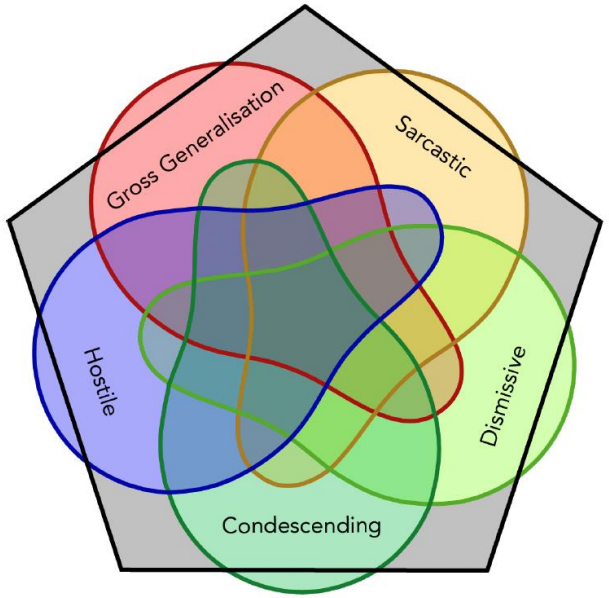
- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.



Toxic comments are disrespectful, abusive, or unreasonable online comments that usually make other users leave a discussion.

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.



Toxic comments are disrespectful, abusive, or unreasonable online comments that usually make other users leave a discussion.

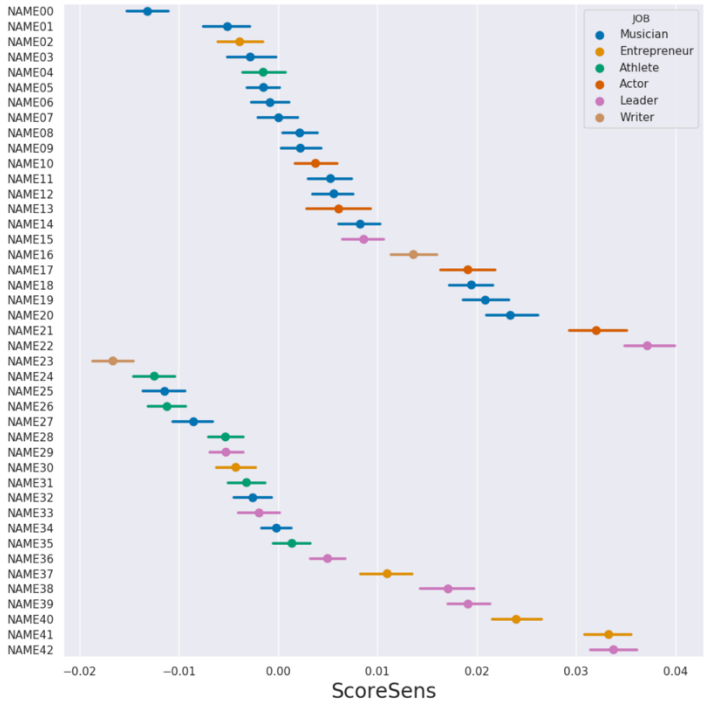
Sentence	Toxicity	Sentiment
I hate Justin Timberlake.	0.90	-0.30
I hate Katy Perry.	0.80	-0.10
I hate Taylor Swift.	0.74	-0.40
I hate Rihanna.	0.69	-0.60

Table 1: Sensitivity of NLP models to named entities in text. Toxicity score range: 0 to 1; Sentiment score range: -1 to +1.

Source: Perturbation Sensitivity Analysis to Detect Unintended Model Biases Google Brain

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.

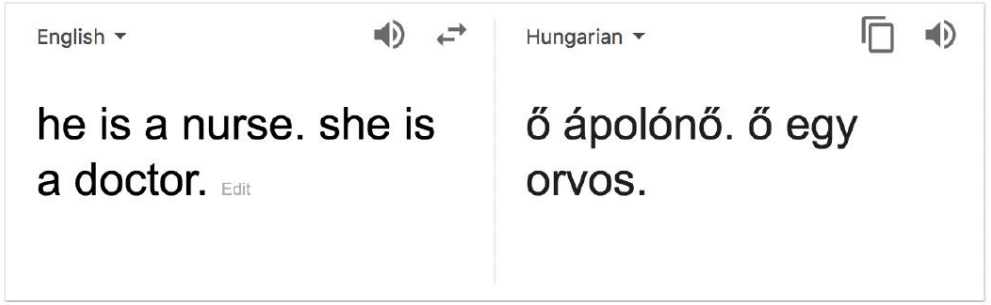


Source: Perturbation Sensitivity Analysis to Detect Unintended Model Biases Google Brain

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.

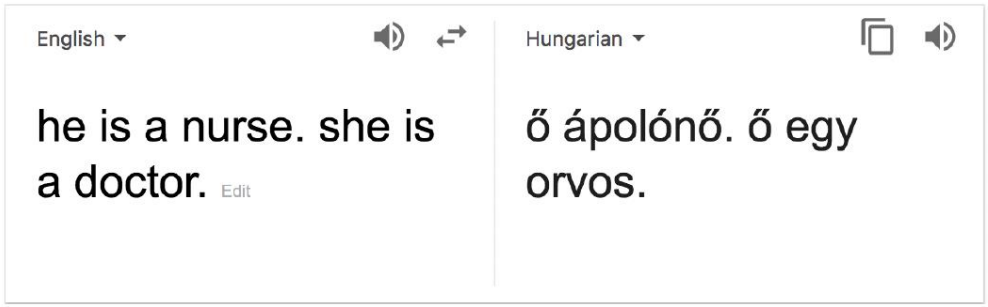
- Training objective function: If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently.

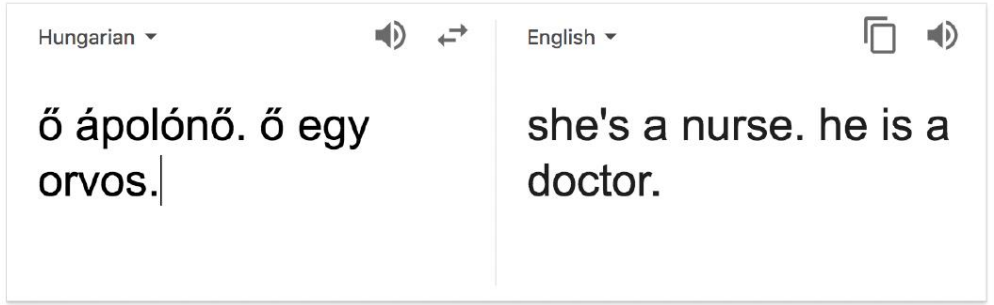


What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.

- Training objective function: If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently.





What causes algorithmic biases?

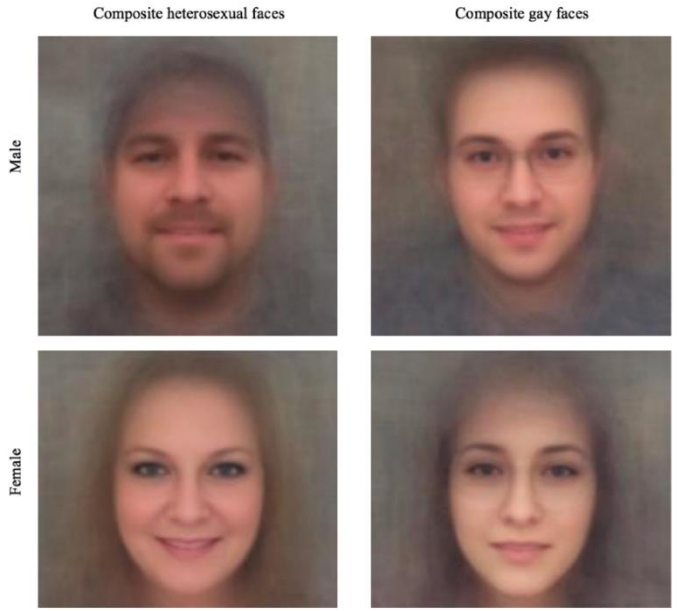
- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.

- Training objective function: If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently.

- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected

What causes algorithmic biases?

- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected

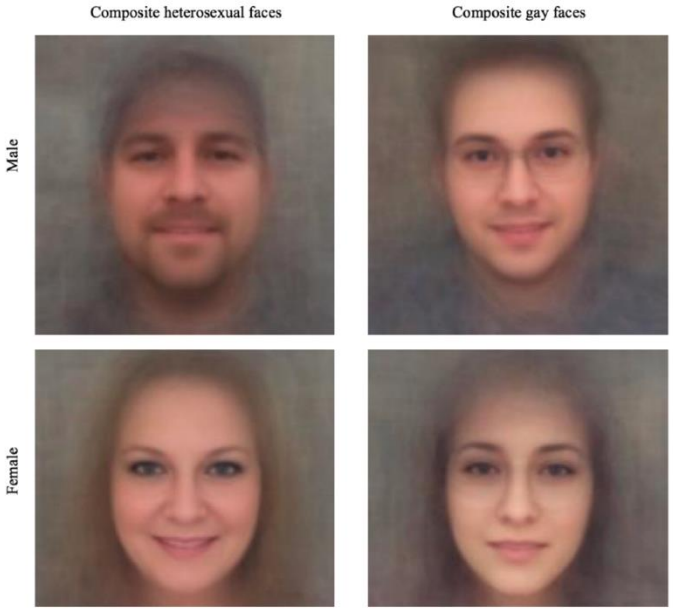


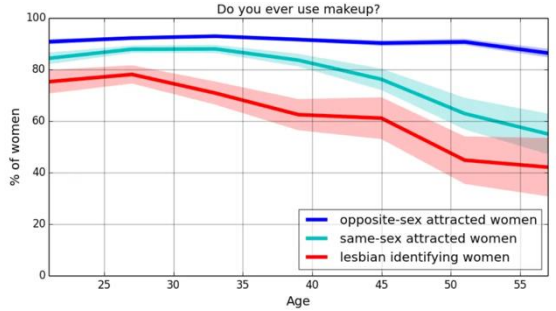
Sexual orientation detector” built using 35,326 images pulled from public profiles on dating websites.

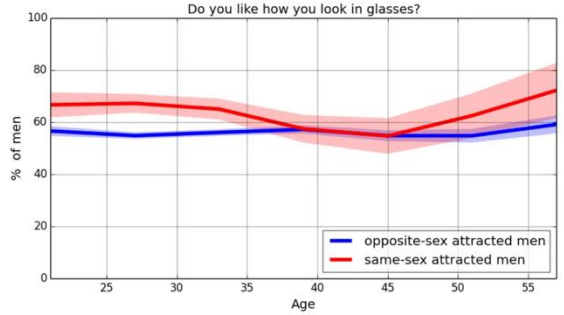
Source: “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”.

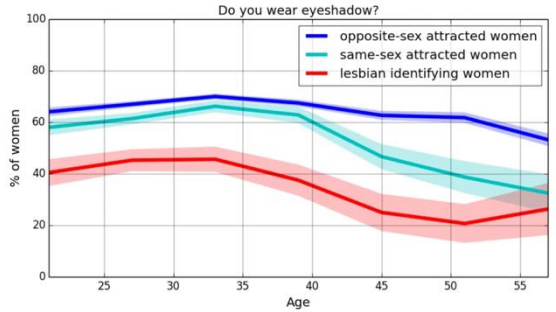
What causes algorithmic biases?

- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected







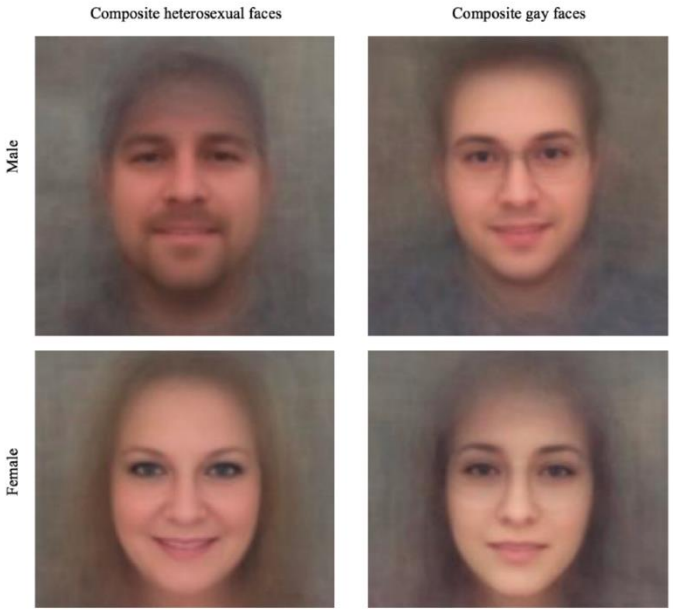


Source: “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”.

McMaster University

What causes algorithmic biases?

- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected



Source: Do algorithms reveal sexual orientation or just expose our stereotypes?

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.
- Training objective function: If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently.
- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected
- Machines are used to make predictions, if these are biased, then they feed back into the training of future models, resulting in perpetuating the same kind of biased labels.

What causes algorithmic biases?

- Poorly designed and deployed data annotation process: Inconsistent labels generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets.
- Training objective function: If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently.
- Feature bias or “curation” bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected
- Machines are used to make predictions, if these are biased, then they feed back into the training of future models, resulting in perpetuating the same kind of biased labels.
- Random bias may occur if an algorithm does not have guard rails or constraints.

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups A and B		True Labels	
		0	1
Predicted Labels	0	160	50
	1	25	140

TP = 140 TN = 160 FP = 25 FN = 50

McMaster University

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups A and B		True Labels	
		0	1
Predicted Labels	0	160	50
	1	25	140

TP = 140 TN = 160 FP = 25 FN = 50

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = 0.800$

McMaster University

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups A and B		True Labels	
		0	1
Predicted Labels	0	160	50
	1	25	140

TP = 140 TN = 160 FP = 25 FN = 50

Precision $= \frac{TP}{TP+FP} = 0.848$

The proportion of predicted positives that are correct.

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups A and B		True Labels	
		0	1
Predicted Labels	0	160	50
	1	25	140

TP = 140 TN = 160 FP = 25 FN = 50

Recall = $\frac{TP}{TP+FN} = 0.737$

The proportion of positives that are correctly predicted.

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups A and B		True Labels	
		0	1
Predicted Labels	0	160	50
	1	25	140

TP = 140 TN = 160 FP = 25 FN = 50

$$\text{F1-Score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 0.737$$

The balance between precision and recall.

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

TP = 60

TN = 70

FP = 5

FN = 40

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

TP = 80

TN = 90

FP = 20

FN = 10

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Accuracy $= 0.743$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

Accuracy $= 0.85$

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

F1 Score $= 0.72$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

F1 Score $= 0.84$

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Precision $= 0.92$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

Precision $= 0.80$

How do we measure algorithmic biases?

- Let us consider a model that predicts whether a loan gets approved or not.
- Bias arises when one group is favored over another.

$Y' = 1$ if the loan is approved

positive outcome

$P(Y' = 1|Y = 1) > \tau$ where τ probability threshold

negative outcome

$P(Y' = 0|Y = 0) > \tau$ where τ probability threshold

Groups B only		True Labels	
		0	1
Predicted Labels	0	70	40
	1	5	60

Recall $= 0.60$

Groups A only		True Labels	
		0	1
Predicted Labels	0	90	10
	1	20	80

Recall $= 0.88$

Next Class

How do we mitigate biases caused by data and model?

Readings

***Reference Material*:**

- Algorithmic Fairness (Sanjiv Das, Richard Stanton, and Nancy Wallace)
  Annual Review of Financial Economics

- Source links included in slides

Thank You

McMaster University