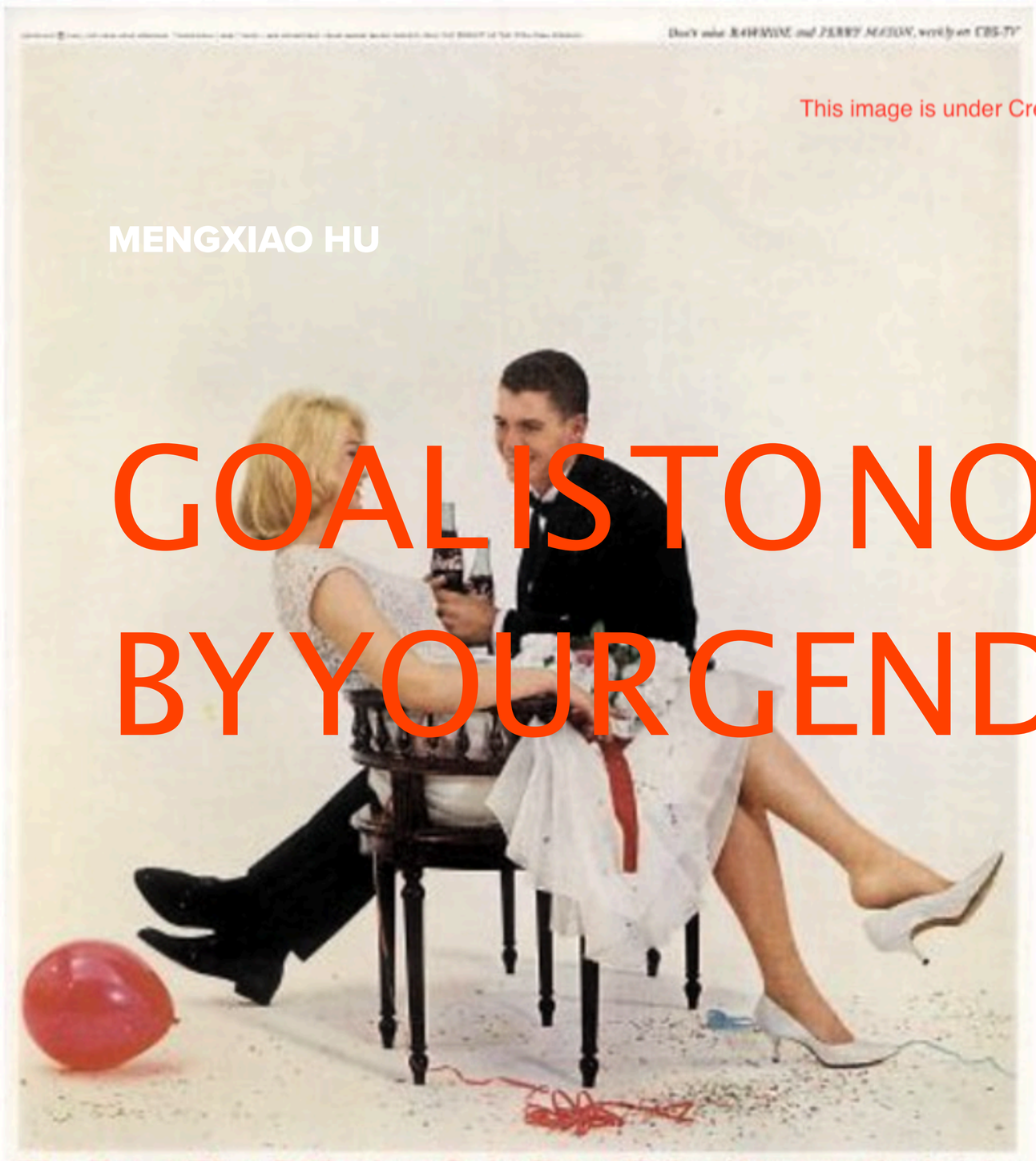
MENGXIAO HU

GOAL IS TO NOT RATE YOU BY YOUR GENDER OR RACE

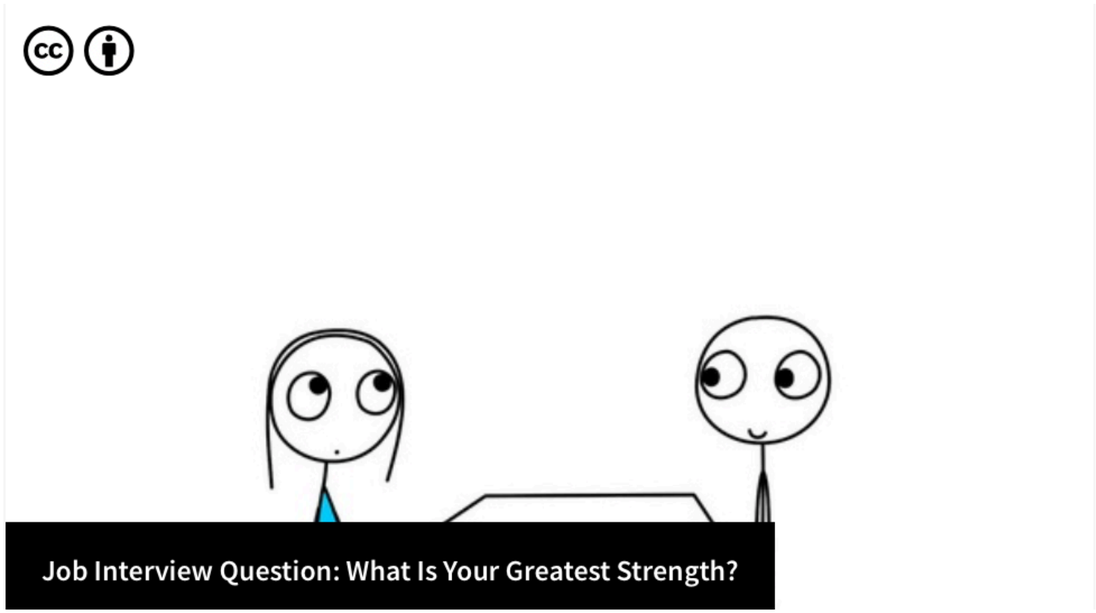


HAVE YOU EVER BEEN ANNOYED BY:

BEING ALLOCATED A DIFFERENT LIMIT OF A CREDIT CARD OR
FINDING YOUR COLLEAGUES GOT HIGHER PAID OR
BEING AVOIDED TO SEAT BY IN A BUS.

Mostly because you are a female or black?

THE BAD THING IS THAT EVEN AI LEARNS TO JUDGE PEOPLE BY THEIR APPEARANCE



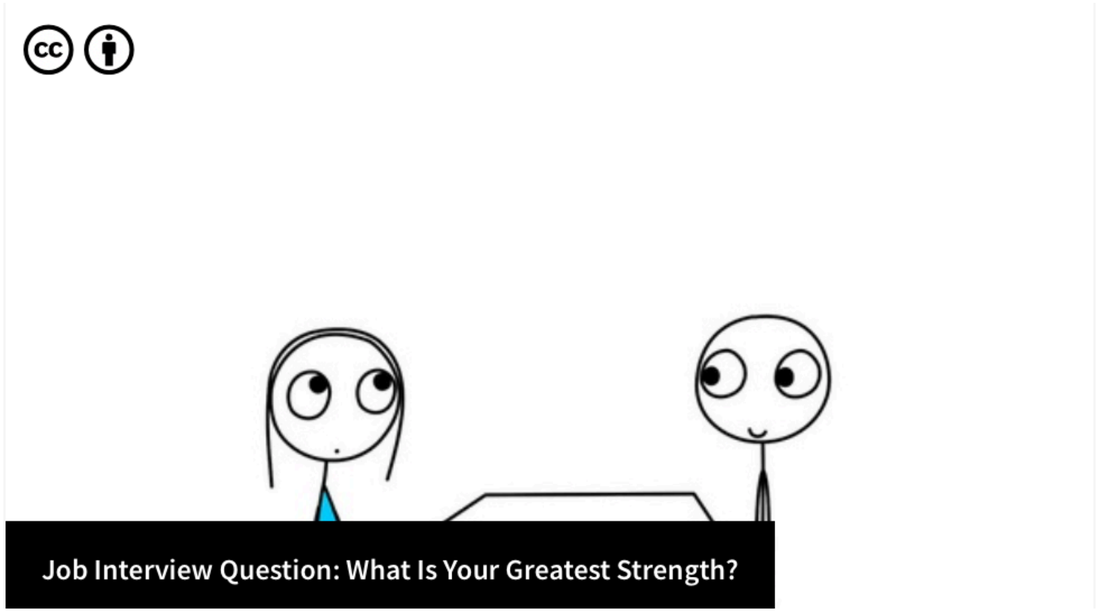
THE BAD THING IS THAT EVEN AI LEARNS TO JUDGE PEOPLE BY THEIR APPEARANCE

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning



THE BAD THING IS THAT EVEN AI LEARNS TO JUDGE PEOPLE BY THEIR APPEARANCE

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning

Healthcare IT News

AI bias may worsen COVID-19 health disparities for people of ...

"If not properly addressed, propagating these biases under the mantle of AI has the potential to exaggerate the health disparities faced by ...

3 weeks ago

THE GOOD THING IS THAT WE ARE SOLVING IT

Axios

Beating AI bias in facial recognition

Identity-verification startup Onfido is training its machine-learning system to reduce the bias that leads AI to make more facial recognition errors ...

3 weeks ago

Datanami

LinkedIn Unveils Open-Source Toolkit for Detecting AI Bias

As AI becomes increasingly integrated in our day-to-day lives, the implications of bias in AI grow more and more worrisome. Training data that ...

2 weeks ago

Toolbox

Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?

Bias in the AI systems stems from training the datasets, where data engineers or data scientists use unconscious cognitive biases starting from ...

1 week ago

THE GOOD THING IS THAT WE ARE SOLVING IT

To let your wages be fairly raised, scholars are contributing to *Fair Machine Learning*.

Axios

Beating AI bias in facial recognition

Identity-verification startup Onfido is training its machine-learning system to reduce the bias that leads AI to make more facial recognition errors ...

3 weeks ago

Datanami

LinkedIn Unveils Open-Source Toolkit for Detecting AI Bias

As AI becomes increasingly integrated in our day-to-day lives, the implications of bias in AI grow more and more worrisome. Training data that ...

2 weeks ago

Toolbox

Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?

Bias in the AI systems stems from training the datasets, where data engineers or data scientists use unconscious cognitive biases starting from ...

1 week ago

THE GOOD THING IS THAT WE ARE SOLVING IT

Axios

Beating AI bias in facial recognition

Identity-verification startup Onfido is training its machine-learning system to reduce the bias that leads AI to make more facial recognition errors ...

3 weeks ago

Datanami

LinkedIn Unveils Open-Source Toolkit for Detecting AI Bias

As AI becomes increasingly integrated in our day-to-day lives, the implications of bias in AI grow more and more worrisome. Training data that ...

2 weeks ago

Toolbox

Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?

Bias in the AI systems stems from training the datasets, where data engineers or data scientists use unconscious cognitive biases starting from ...

1 week ago

To let your wages be fairly raised, scholars are contributing to *Fair Machine Learning*.

For example, in Zhao's work:

THE GOOD THING IS THAT WE ARE SOLVING IT

Axios

Beating AI bias in facial recognition

Identity-verification startup Onfido is training its machine-learning system to reduce the bias that leads AI to make more facial recognition errors ...

3 weeks ago

Datanami

LinkedIn Unveils Open-Source Toolkit for Detecting AI Bias

As AI becomes increasingly integrated in our day-to-day lives, the implications of bias in AI grow more and more worrisome. Training data that ...

2 weeks ago

Toolbox

Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?

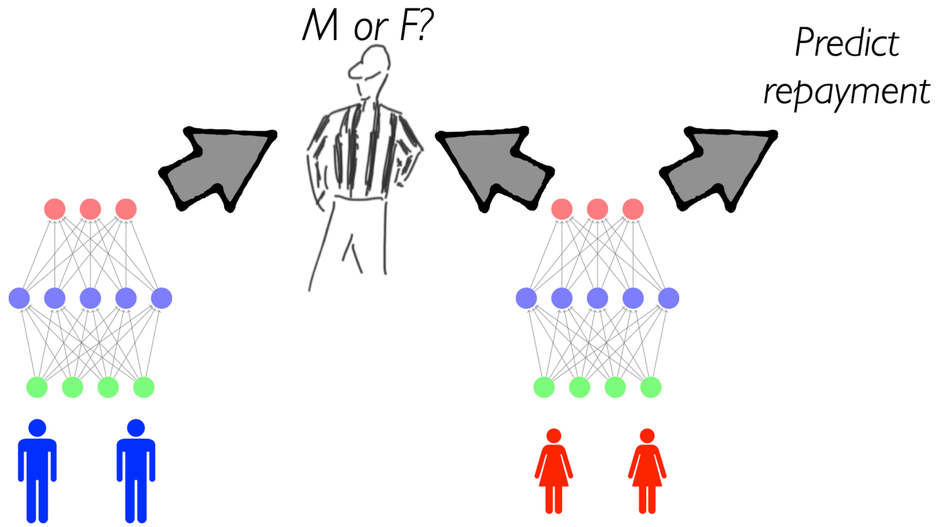
Bias in the AI systems stems from training the datasets, where data engineers or data scientists use unconscious cognitive biases starting from ...

1 week ago

To let your wages be fairly raised, scholars are contributing to *Fair Machine Learning*.

For example, in Zhao's work:

https://hanzhaoml.github.io/papers/ICLR2020/cfair_slides.pdf



They gave many forms of fairness...

https://hanzhaoml.github.io/papers/ICLR2020/cfair_slides.pdf

Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

[Verma et al. 18]

We have many works to do:

- When do we use them?
- Can we generalize them into fewer ones?
- How to design algorithms with them?

In Zhao's work:

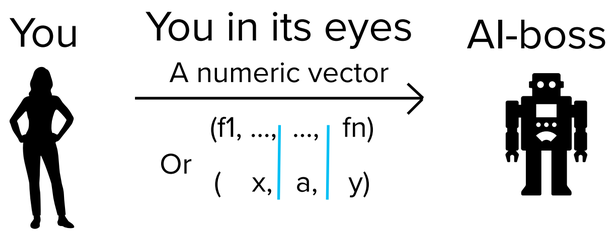
- Are they consistent?
- At least some of them?

IN REPRESENTATION LEARNING CONTEXT,

- **He found two definitions of fairness are not inconsistent;**

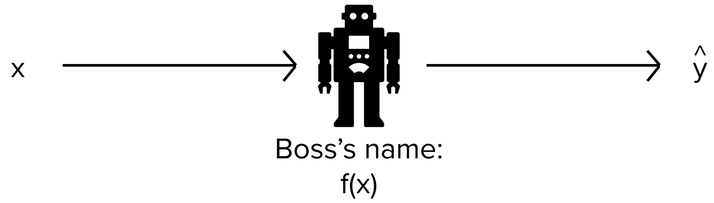
Because the objective function he proposed simultaneously ensured them.

IT CAN REDUCE THE BIAS OF AN AI-BOSS'S DECISIONS ON YOUR WAGES, BUT HOW...



a: gender
y: biased wage
x: other features

For fairness, 'a' is hidden for decisions



Suppose: $(X, A, Y) \sim \mathcal{D}$

- **Equalized odds:** $\hat{Y} \perp A \mid Y$
- **Accuracy parity:** $\mathrm{err}(\hat{Y}) \perp A$

- He found these two definitions of fairness are not inconsistent;

SO, WE CAN NOT TRADE FAIRNESS FOR ACCURACY!

IT IS IMPOSSIBLE IF WE USE ANOTHER FORM OF FAIRNESS

- Demographic parity*: $\hat{Y} \perp A$

BECAUSE DP IS NOT CONSISTENT WITH AP

$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$ [Zhao and Gordon, NeurIPS 19]

$\Delta_{\mathrm{BR}} := |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$

*Also called statistical parity.

HE PROPOSED A NEW METRIC TO SIMULTANEOUSLY ENSURE EO AND AP!

$$\mathrm{BER}_{\mathcal{D}}(\widehat{Y} \mid\mid Y) := \mathcal{D}(\widehat{Y} = 0 \mid Y = 1) + \mathcal{D}(\widehat{Y} = 1 \mid Y = 0)$$

He showed that in some cases, EO is satisfied, the BER is the upper bound of Type-I and Type-II errors:

$$\mathrm{Err}_{\mathcal{D}_0}(\widehat{Y}) + \mathrm{Err}_{\mathcal{D}_1}(\widehat{Y}) \leq 2\mathrm{BER}_{\mathcal{D}}(\widehat{Y} \mid\mid Y)$$

That means we can minimize the BER to ensure the accuracy!

WHAT'S THE CASE?

$$\Pr_{A=0}(\widehat{Y} = 1 \mid Y = 0) \approx \Pr_{A=1}(\widehat{Y} = 1 \mid Y = 0)$$

And,

$$\Pr_{A=0}(\widehat{Y} = 0 \mid Y = 1) \approx \Pr_{A=1}(\widehat{Y} = 0 \mid Y = 1)$$

THE SECOND TERM IS ENCOURAGING THIS CASE:

$$\min_{h,g} \max_{h',h''} \quad \mathrm{BER}_{\mathcal{D}}(h(g(X)) \mid\mid Y) - \lambda \left(\mathrm{BER}_{\mathcal{D}^0}(h'(g(X)) \mid\mid A) + \mathrm{BER}_{\mathcal{D}^1}(h''(g(X)) \mid\mid A)\right)$$


THAT MEANS IT ENCOURAGES TO
SIMULTANEOUSLY MINIMIZE THE ERROR RATE
AND EO!

FAIR AND RIGHT WAGES ARE ENSURED