

AI + BIASES



INTRODUCTION

I'm Anusree..
AI Engineer at Innovation Incubator
Advisory

I work from recommendation systems
to defect detection systems



Let' s talk about human
biases







Bias in Image Classification



Ground Truth: Bride



CNN for image classification.



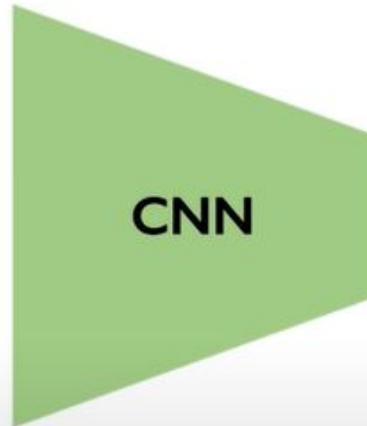
Predicted Classes

Bride
Dress
Ceremony
Woman
Wedding

Bias in Image Classification



Ground Truth: Bride



CNN

CNN for image
classification.



Predicted Classes

Clothing
Event
Costume
Red
Performance art

AI BIAS IS AN ANOMALY IN THE OUTPUT OF MACHINE LEARNING ALGORITHMS. THESE COULD BE DUE TO THE PREJUDICED ASSUMPTIONS MADE DURING THE ALGORITHM DEVELOPMENT PROCESS OR PREJUDICES IN THE TRAINING DATA.

AI Bias



SOME EXAMPLES



US HEALTHCARE

Race as a variable



COMPAS

Correctional Offender
Management Profiling for
Alternative Sanctions



AMAZON HIRING ALGO

hiring employees
was found to be
biased against
women



FB ADS

Gender bias

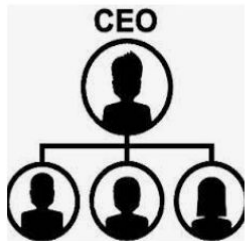
forbes.com

execed.economist.com

inc.com

ceotodaymagazine.com

europeanceo.com



What is CEO?
computerhope.com



CEO vs. CIO vs. COO vs. Other Executives
digital-adoption.com



Why You Need To Be The CEO Of Your ...
forbes.com



CEO Job Description
betterteam.com



Role Of Personality In CEO Leadership ...
chiefexecutive.net



LinkedIn CEO Jeff Weiner steps d...
fortune.com



hot CEO* romantic novels
managementtoday.co.uk



CEO Survival Guide: Leading with ...
entrepreneur.com



Why CEOs Actually Deserve Their ...
time.com



Best CEOs 2019 List ...
elcompanies.com



How to Become a CEO: Definition Ste...
online.maryville.edu



How CEOs can help lead te...
mckinsey.com



Related searches



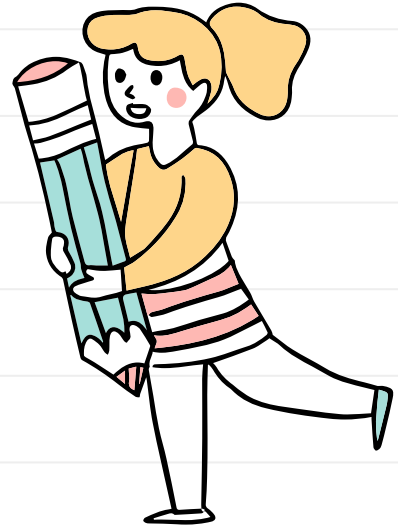
christmas message to employees from ceo



managing director cv



LEARN THE ROOT CAUSE



SOURCE OF BIAS

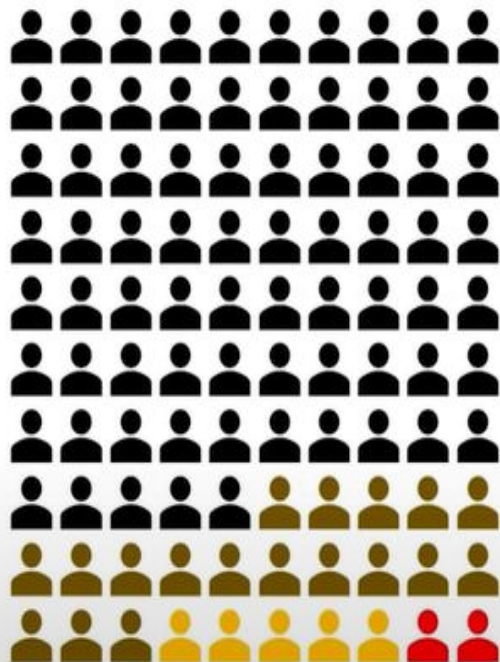
TRAINING DATA

Negative legacy
Under-estimation
Oversampling
Data Augmentation

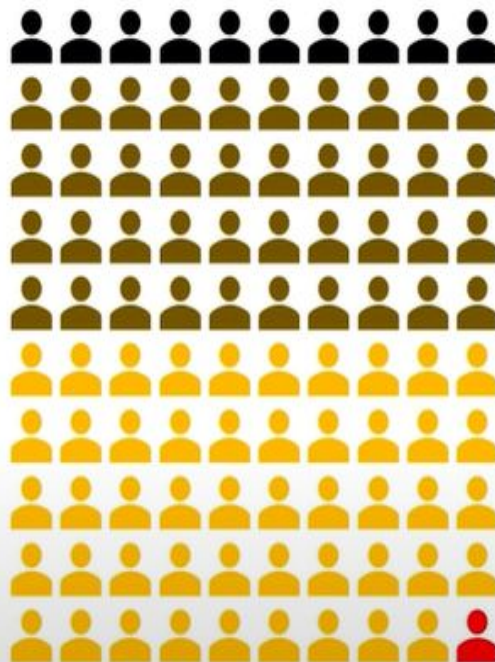
ALGORITHMS

Proxy discrimination
Protected attributes
Counterfactual fairness

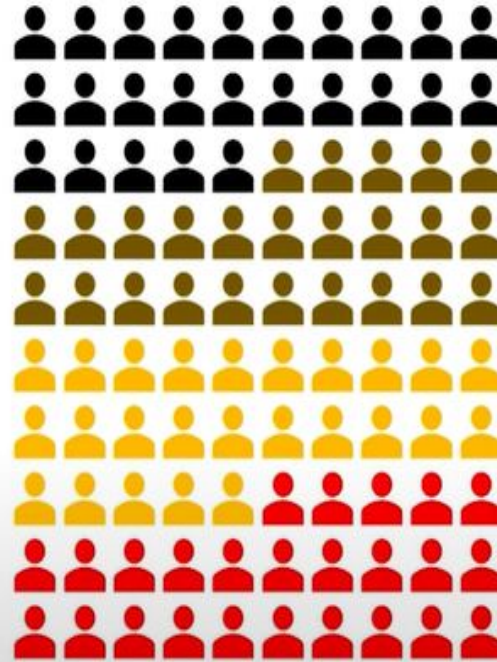
Real World



"Gold-Standard" Dataset



Balanced Dataset



Black Hair



Brown Hair



Blonde Hair



Red Hair

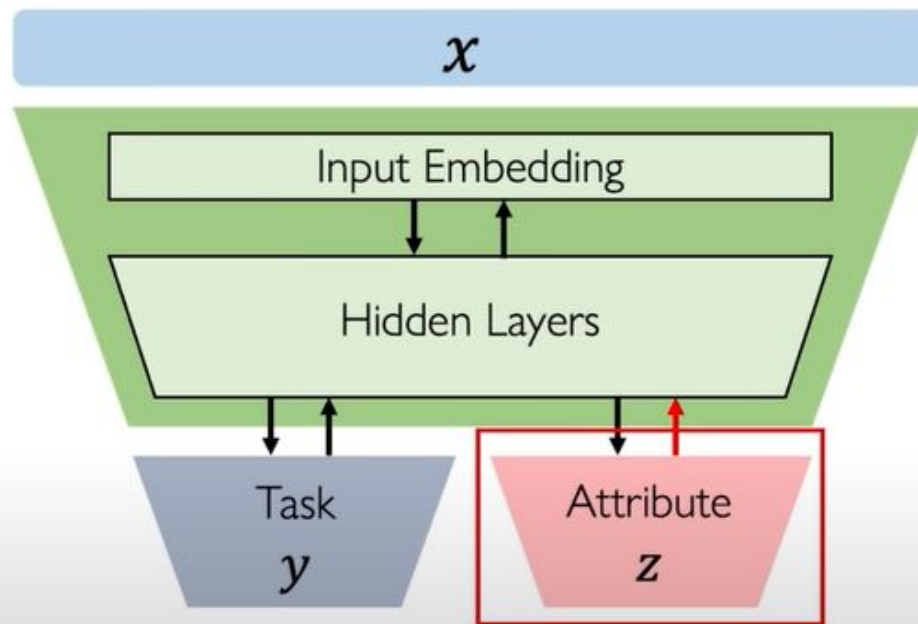
Application to Language Modeling

Task: language model to complete analogies

He is to **she**, as **doctor** is to ?

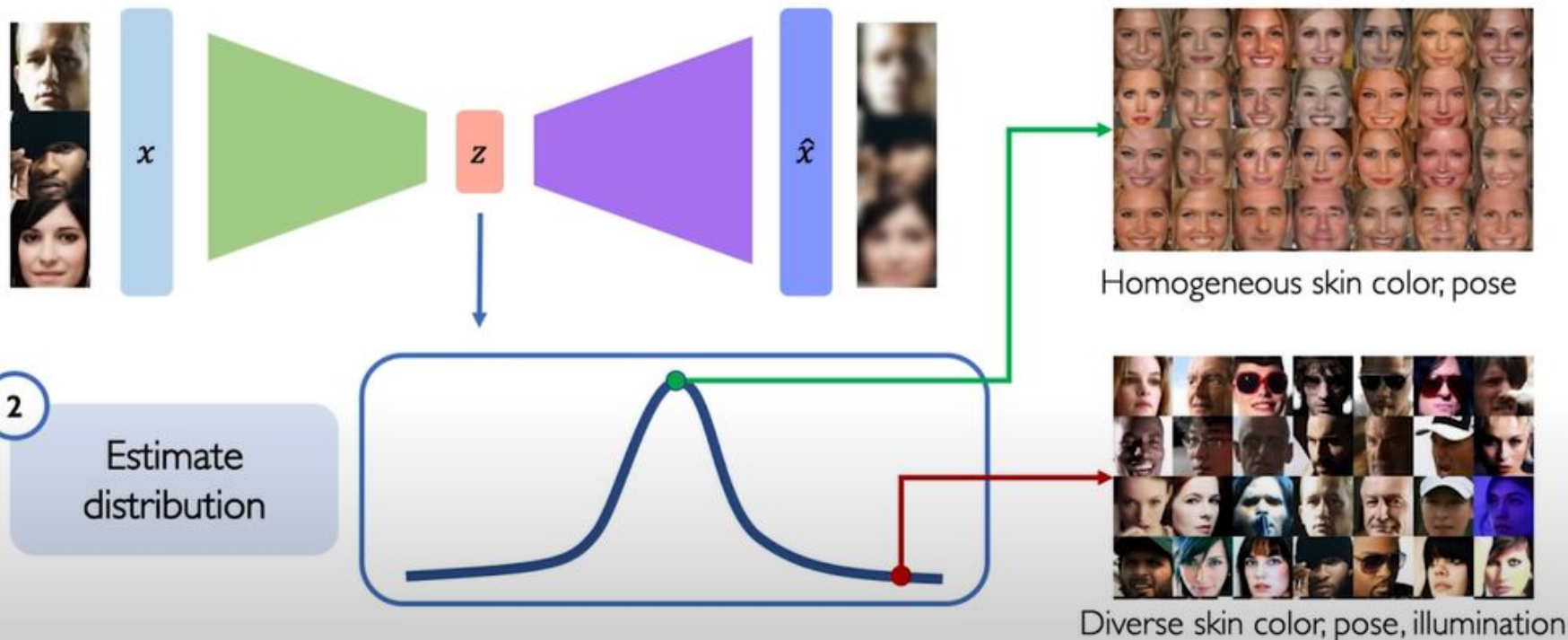
biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Sensitive attribute: Gender



Jointly predict output label y and sensitive attribute z to remove from decision

Mitigating Bias through Learned Latent Structure



Using Latent Variables for Automated Debiasing

Approximate the distribution of the latent space with a joint histogram over the latent variables:

$$\underbrace{\hat{Q}(z|X)}_{\text{Estimated joint distribution}} \propto \prod_i \underbrace{\hat{Q}_i(z_i|X)}_{\text{Histogram for each latent variable } z_i}$$

Independence to approximate

Define **adjusted probability** for sampling a particular datapoint x during training:

$$\underbrace{W(z(x)|X)}_{\text{Probability of selecting datapoint}} \propto \prod_i \frac{1}{\underbrace{\hat{Q}_i(z_i(x)|X)}_{\text{Histogram for each latent variable } z_i} + \underbrace{\alpha}_{\text{Debiasing parameter}}}$$

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider

1



Be aware of contexts in which AI can help correct for bias and those in which there is high risk for AI to exacerbate bias

2



Establish processes and practices to test for and mitigate bias in AI systems

3



Engage in fact-based conversations about potential biases in human decisions

4



Fully explore how humans and machines can best work together

5



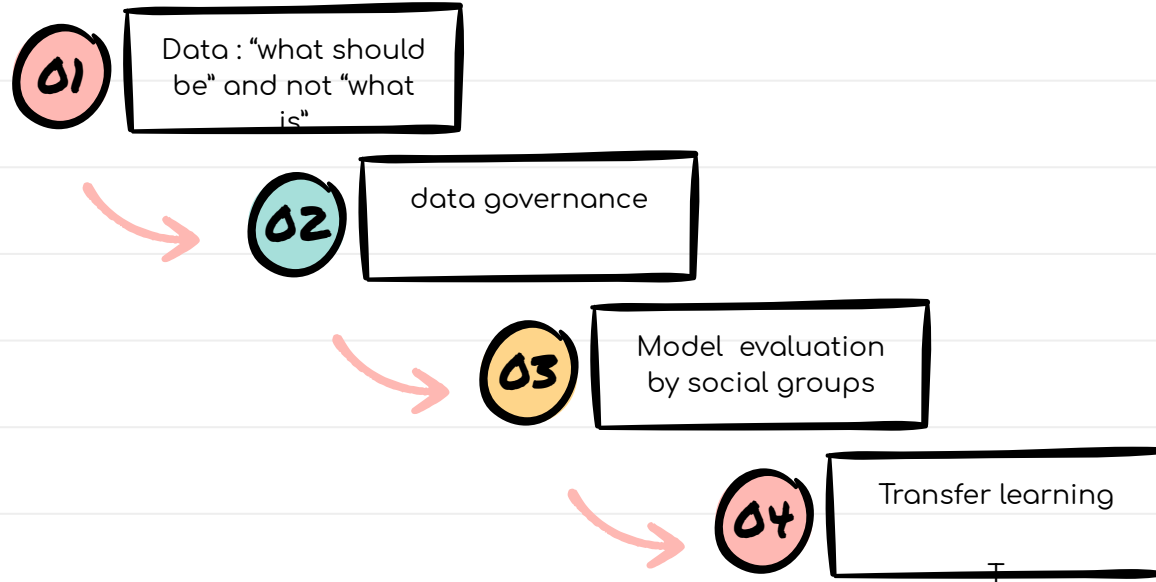
Invest more in bias research, make more data available for research (while respecting privacy), and adopt a multidisciplinary approach

6



Invest more in diversifying the AI field itself

HOW TO TACKLE IT



TOOLS

AI Fairness 360 - for
binary classification



TO



Google's What-If
Tool


REDUCE



IBM Watson
OpenScale

BIAS

Visualize

☒ Datapoints ☐ Partial dependence plots☐ Show nearest counterfactual datapoint ☒ L1 ☐ L2  

Edit

      Search features

Select a datapoint to begin exploring model behavior for your selection.

Edit and Infer: Edit your datapoint here and run inference in the Infer table to see differences in model behavior.

Visualize: Switch between visualizing datapoints and exploring partial dependence plots to gain insights into your model's behavior. Explore counterfactuals or see how similar (or different) the rest of your dataset is from your selection.

Infer



Binning | X-Axis

(none)



Binning | Y-Axis

(none)



Color By

Inference label 1



Label By

(default)



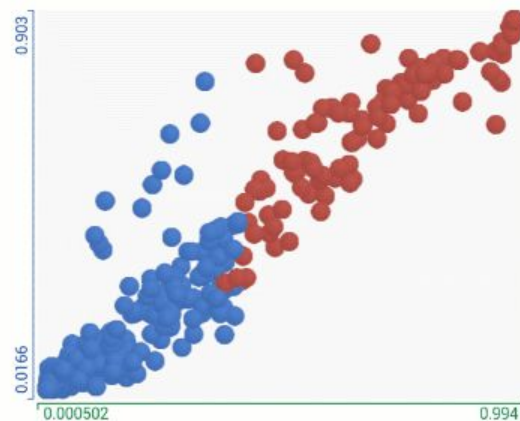
Scatter | X-Axis

Inference score 1



Scatter | Y-Axis

Inference score 2



Legend



Colors

by Inference label 1

- <=50k
- >50k

THANKS...