Bias, Fairness, and Ethics

Terminology

- Bias (vernacular): prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair. (OED)
- Bias (statistics): the difference between the expected value of an estimator and the true population mean.

Algorithmic Bias

- “I’m training a network to classify images”
- “I’m training a network to generate text”
- We are *always* training the network to replicate data
- Biases in the data are encoded in the network

Algorithmic Bias Examples

- Recidivism Prediction[1]

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

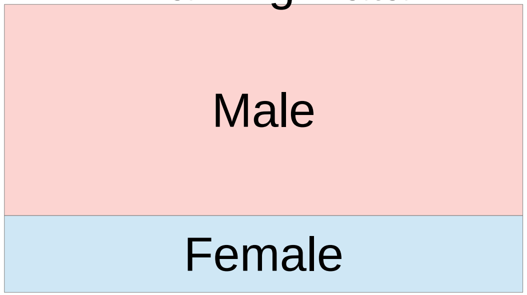
- Recruiting[2]

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

Training Data



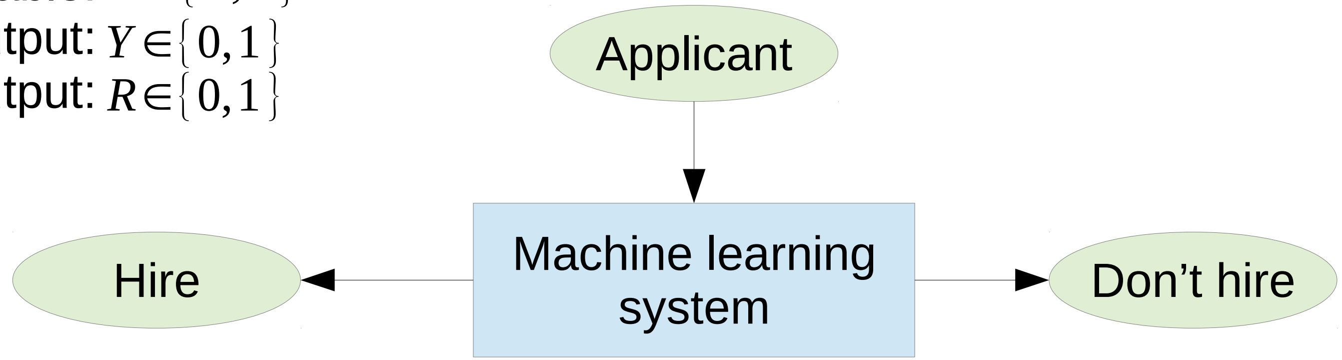
[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." May 23, 2016. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Jeffrey Dastin. "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women." Oct. 10, 2018. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Fairness

Binary Classification

Sensitive Variable: $A \in \{a, b\}$
True output: $Y \in \{0, 1\}$
Predicted output: $R \in \{0, 1\}$

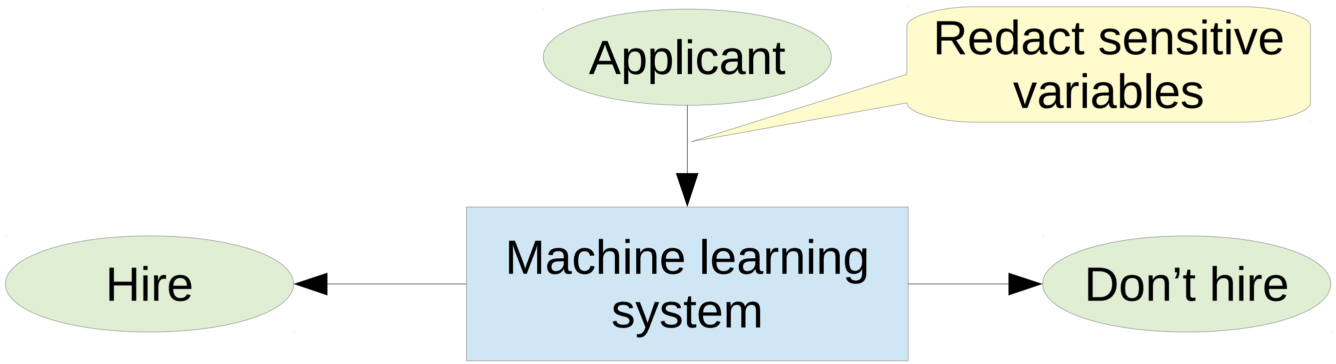


Sensitive variables

E.g., protected classes in hiring: race, religion, gender, disability, veteran status, etc.

Much of this presentations discussion of fairness is based on “A Tutorial on Fairness in Machine Learning” by Ziyuan Zhong. Published Oct 21, 2018, accessed April 15, 2022.
https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Unawareness



- Both of the previous examples were “unaware”
- Correlated features
  - E.g., zip code, name

Demographic Parity

$$P(R=1|A=a)=P(R=1|A=b)$$

- ✔ Legally motivated (four-fifths rule)
- ✗ Lazy solution: predict correctly on one group and randomly on the other

Predictive Parity

$$P(Y=1|R=1, A=a) = P(Y=1|R=1, A=b)$$

- ✔ Encourages good predictions
- ✗ Can be reflective of true societal bias

Impossibility Theorem

Demographic parity does not hold

$$\text{If } \neg(A \perp Y) \text{ and } A \perp Y | R \text{, then } \neg(A \perp R)$$

If the true outcome depends on the sensitive variable then either

- The prediction depends on the sensitive variable, or
- The true outcome conditioned on the prediction depends on the sensitive variable

Predictive parity does not hold

For a deeper dive on fairness, see “Fairness and Machine Learning” by Solon Barocas, Moritz Hardt, and Arvind Narayanan. Available at https://fairmlbook.org/index.html.

Ethics

Who is developing models and why?

Deliberate Misuse

Synthesizing Obama: Learning Lip Sync from Audio

SIGGRAPH 2017

Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman

Output Obama Video

Incentives and Limitations

- Many learning algorithms are designed to make money

Auditing Radicalization Pathways on YouTube

Manoel Horta Ribeiro*
EPFL
manoel.hortaribeiro@epfl.ch

Raphael Ottoni
UFMG
rapha@dcc.ufmg.br

Robert West
EPFL
robert.west@epfl.ch

Virgílio A. F. Almeida
UFMG, Berkman Klein Center
virgilio@dcc.ufmg.br

Wagner Meira Jr.
UFMG
meira@dcc.ufmg.br

(FAccT 2020)

- Who can train models?

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
. . .								
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

>30 years on an RTX 3090 Ti

~7 months on TACC's Stampede2

Tom B. Brown et al. "Language Models are Few-Shot Learners." NeurIPS 2020.

What Can You Do?

- As a user: normal media literacy
- Understand your data
- Analyze your models
- Actively work on bias/fairness