The Alan Turing Institute

Introduction to Fairness

Dr. Debashish Das

What We Will Cover?

- Conception of Fairness in AI
- Fairness Issues Need to be Addressed by AI
- Individual & Group Fairness, and Tension between Each
- Fairness, Privacy, and Transparency
- Example of Unfairness in AI
- Unfairness in Model: Bias Amplification
- Fairness in Automated Decisions
- Discrimination
- Unfairness in Machine Learning
- Subtler Bias
- Data Issues
- Unfairness in Data Collection
- Unfairness in the World
- Demographic Disparities
- Model Issues
- Fair Classification
- How do we make a fair model?
- Add Fairness Constraints
- Machine Learning Pipeline
- Fairness in ML : Goals
- Stages of ML System
- Further Thoughts
- Summary

What We Will Cover in Module 1?

- Introduction to Fairness
- Fairness of Data, Algorithms & Models
- Individual Fairness, Group Fairness, and the tension between them
- Fairness, Privacy, and Transparency by Design in AI/ML Systems

Conception of Fairness in AI

- Fairness and Equality are critical aspects in the various domains like Governments, Credit, Advertising, Recruiting etc.
- With the widespread use of artificial intelligence (AI) systems Fairness has gained significant importance
- AI systems can be used in many sensitive environments to make even life-changing decisions
- It is vital to ensure that the decisions do not reflect discriminatory behaviour toward certain groups or populations
- Fairness in AI refers to the process of correcting and eliminating algorithmic bias (race, ethnicity, gender, sexual orientation, disability, and class) from AI models

Fairness Issues Need to be Addressed by AI

– Data (input)

--Inaccurate or insufficient data
--Underrepresentation

– Models (output)

--Discriminatory treatment of subpopulations
--Build or “post-process” models with subpopulation guarantees
--Not fully causally related variable that are correlated to the protected class

– Algorithms (process)

--Learning algorithm generating data through its decisions
--Lack of clear train/test division

Individual & Group Fairness, and Tension between Each

- At an individual level, fairness can be defined as similar individuals being treated similarly
- At a group level, a fair outcome demands the existence of parity between different protected groups, such as those defined by gender or race
- These measures can give rise to conflicts in situations where, in an attempt to satisfy group fairness, individuals who are similar with respect to the classification task, receive different outcomes
- In practice, the conflict may be resolved by a nuanced consideration of the sources of `unfairness' in a particular deployment context, and the carefully justified application of measures to mitigate it

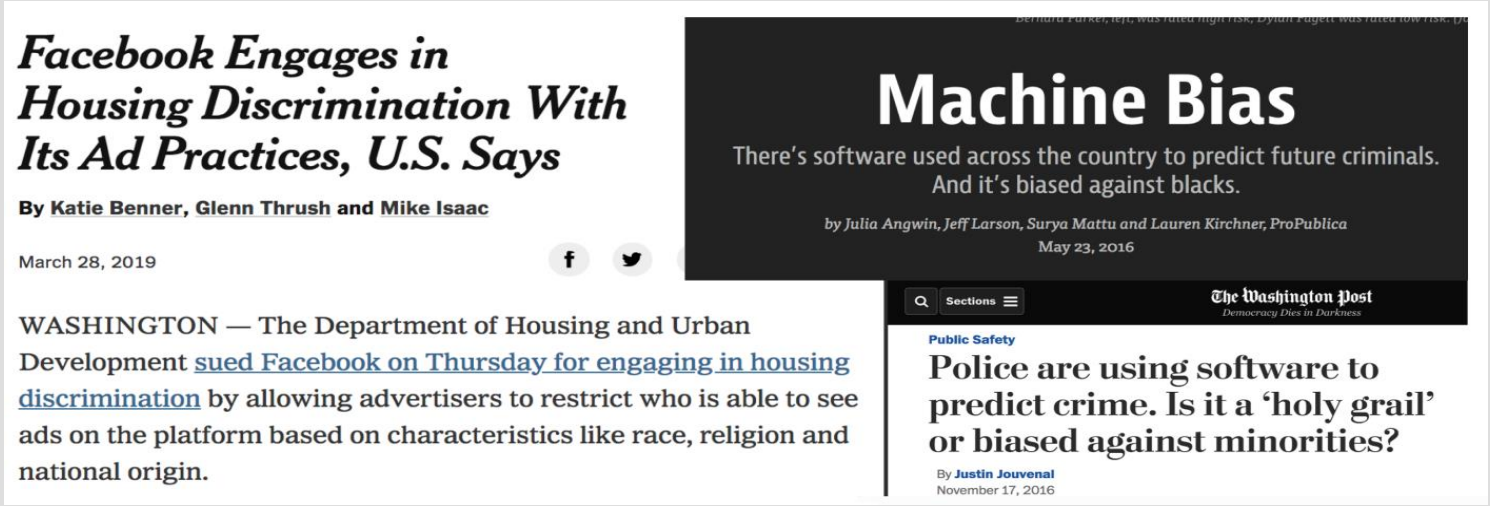
Fairness, Privacy, and Transparency

- First, can we trust technology to be fair, especially given that the data on which the technology is based are biased in various ways?
- Second, whom can we blame if the technology goes wrong, as it inevitably will on occasion?
- Finally, does it matter if we do not know how an algorithm works or, relatedly, cannot understand how it reached its decision?
- Although, the above are serious concerns, they are not irresolvable

Example of Unfairness in AI

– Why was I not shown the Advertisement?

Example of Unfairness in AI (Cont.)



Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Sections

The Washington Post

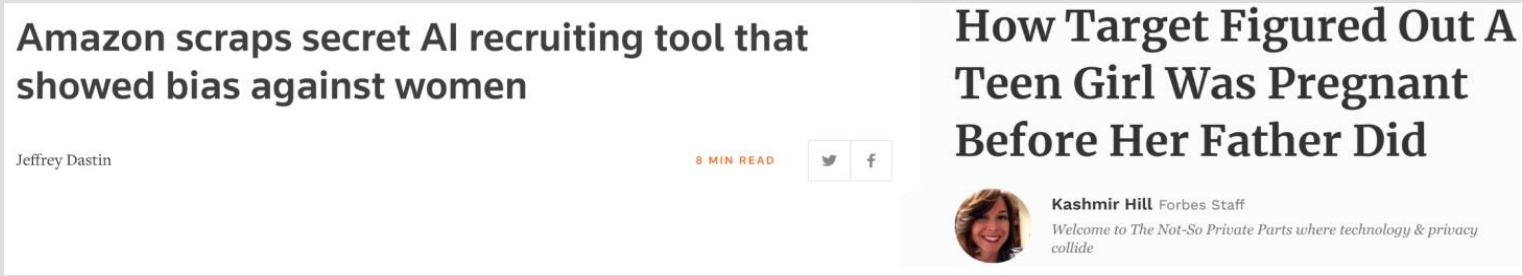
Democracy Dies in Darkness

Public Safety

Police are using software to predict crime. Is it a 'holy grail' or biased against minorities?

By Justin Jouvenal

November 17, 2016



Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

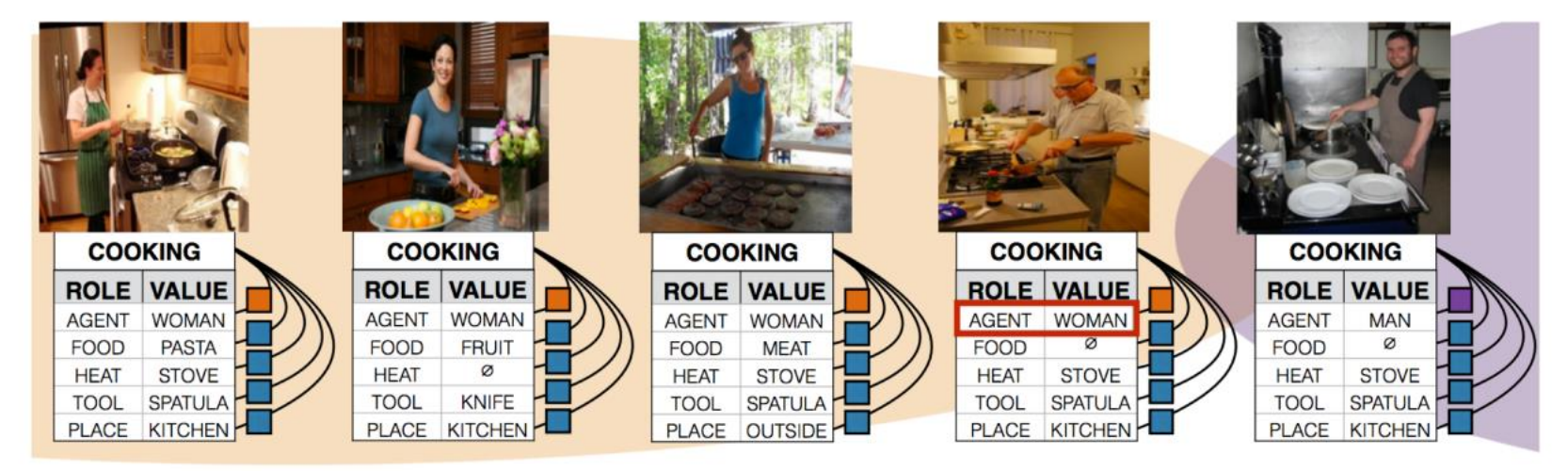
8 MIN READ

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

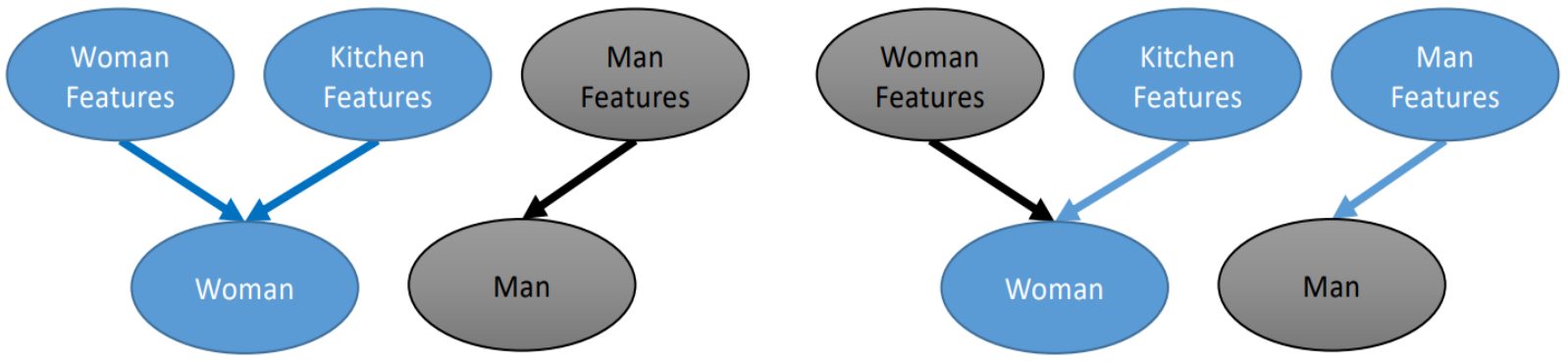
Unfairness in Model: Bias Amplification



Class	Man	Woman
Data prior	33%	67%
Pred. prior	16%	84%

Unfairness in Model: Bias Amplification

– A model demonstrates bias amplification if the prior distribution of the model's predictions does not match that of the data. Here, the aim of the model is to avoid creating or exaggerating disparities in the training data



Fairness in Automated Decisions

- Algorithmic unfairness: Algorithms are pervasive, high-stakes, high-impact
- Need more than just “accuracy”
- What’s changed? Pervasiveness of ML & Attention to demographic Criteria

Discrimination

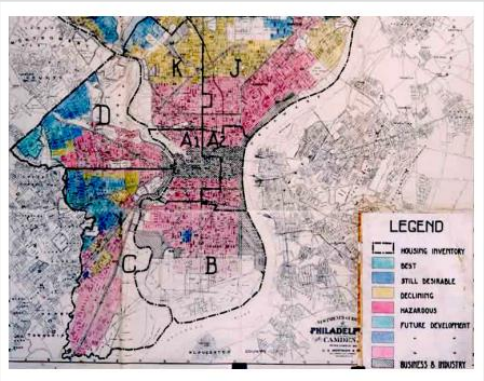
- It is the concern
- Population includes minorities
  - --Ethnic, Religious, Medical, Geographic
- Protected by Law, Policy, Ethics
- If we cannot completely control our data, can we regulate how it is used, how decisions are made based on it?

Forms of Discrimination

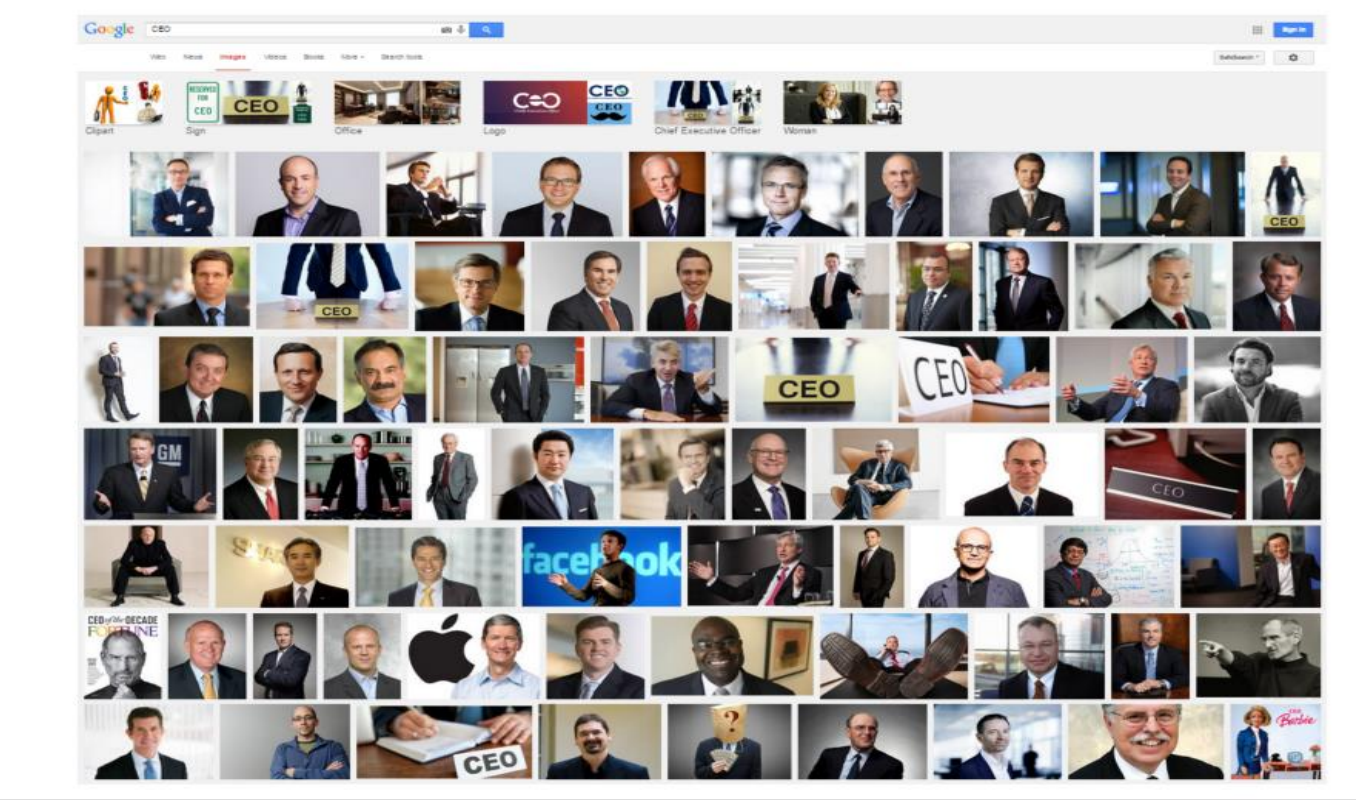
– Steering minorities into higher rates (advertising)

– Redlining: deny service, change rates based on area

– Self-fulfilling prophecy: select less qualified to “justify” future discrimination



Subtler Bias



Subtler Bias

Data Issues

- Basic data issues: imbalanced, impoverished; noisy
- Measurement involves subjective choices, and technical difficulties
- Example: “Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago.” NYT, 2017

  --%age change 1980-2015 in black, Hispanic, Asian, white, multiracial students
- Target variable / labels:

  -- what is “creditworthiness”; “good employee”; “attractive”

  -- objective measures may be biased too

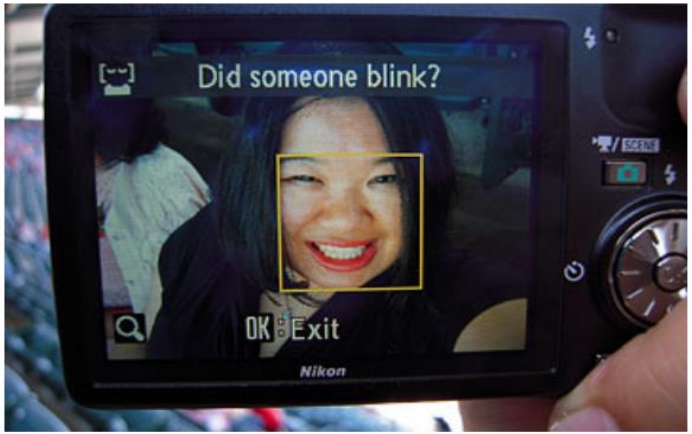
  -- classification schemes may rely on historical taxonomies
- Even images not unbiased

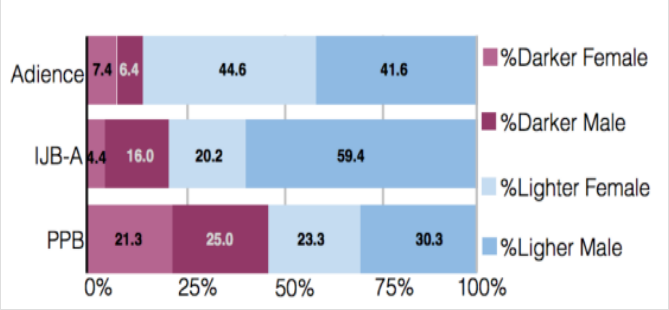
  -- color balance, dynamic range settings distribution of subjects not matching in training/testing

Unfairness in Data Collection

– Nikon blink-recognition always thinks Asian faces are blinking

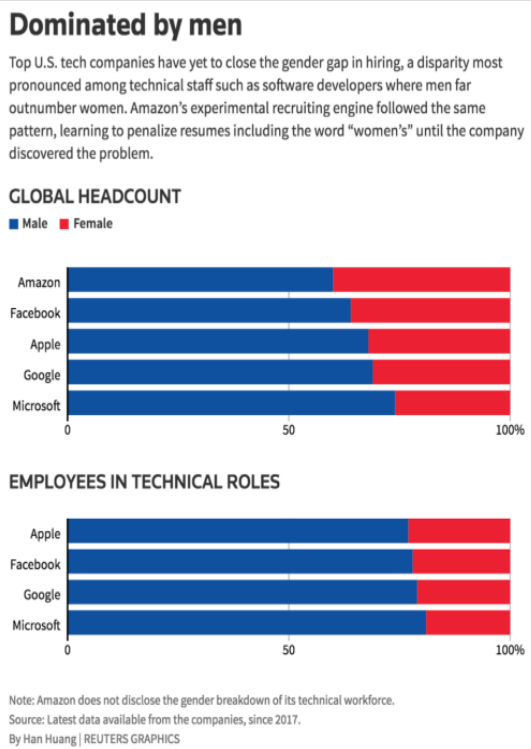
--While it's not certain why the problem exists (Nikon has not given concrete explanation), it's feasible that it is due to unbalanced training data





The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Unfairness in the World: Amazon Recruitment



BUSINESS NEWS

OCTOBER 9, 2018 / 11:12 PM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

- Amazon has a history of hiring predominantly men
- Amazon recruitment tool learned to penalize women's applications to match the distribution in the biased training data
- penalize the word "women" e.g. "women's soccer coach" etc
- favor words more often used in men's applications, eg "execute"

Demographic Disparities

- Most ethical issues arise when data concerns people
- Training data tends to encode demographic disparities in our society -- can perpetuate stereotypes
- Some occupations have stark gender imbalance -- why?
- But not all applications involve people. Or do they? examples: StreetBump; Automated Essay Scoring; Zillow

Model Issues

- Models can faithfully reflect disparities in data, often including stereotypes – why?
- Some patterns we think are good features for classification, others are not: how to tell them apart?

- Can also introduce disparities when none exist – not enough data
- Need to train based on something other than just overall accuracy

Fair Classification

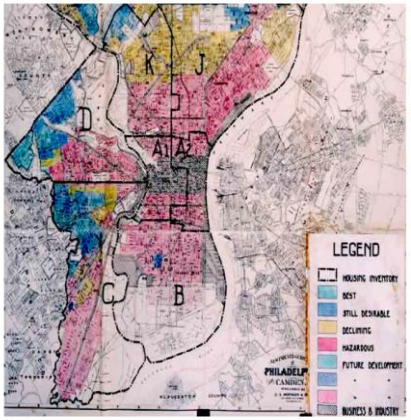
- Explosion of fairness research over the years
- Fair classification is the most common setup, involving:
  - -- X, some data
  - -- Y, a label to predict
  - -- $\hat{Y}$, the model prediction
  - -- A, a sensitive attribute (race, gender, age, socioeconomic status)
- We want to learn a classifier that is:
  - -- accurate
  - --fair with respect to A

How do we make a fair model?

- Try to ensure our model doesn’t augment bias
- Train with a balanced dataset
- Audit your model
- Ensure some constraint, e.g. demographic parity
- Don’t use protected attribute

What happens if we take out the protected attribute?

- Neighborhoods in America are largely racially segregated
- A race-blind model could still act in a discriminatory manner by using zipcode to e.g. deny a loan
- Even unintentional discrimination can occur in this way, given a biased prior



Some Amazon Prime services seem to exclude many predominantly black zip codes

Rafi Letzter Apr. 21, 2016, 12:36 PM

In fact, taking out the protected attribute can be detrimental to fairness goals

- Imagine an AI for hiring new employees has two features: gender and experience.

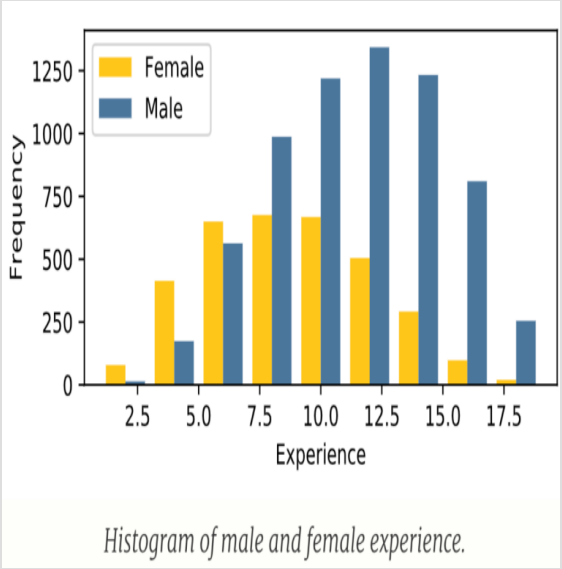
  -- The model hires 27% women, despite their being 44% of the applicant pool.

- In an effort to make a fair model, you take out the gender variable and only use experience

  -- You find your model now hires 17% women.

In fact, taking out the protected attribute can be detrimental to fairness goals (Cont.)

- Perhaps in reality, people with over ten years of experience are equally qualified for the job
- Women have to take off more time due to extenuating circumstances (needing to take family or child leave, etc)
- Removing the gender feature from the model makes it impossible for the model to compensate



Histogram of male and female experience.

Add Fairness Constraints

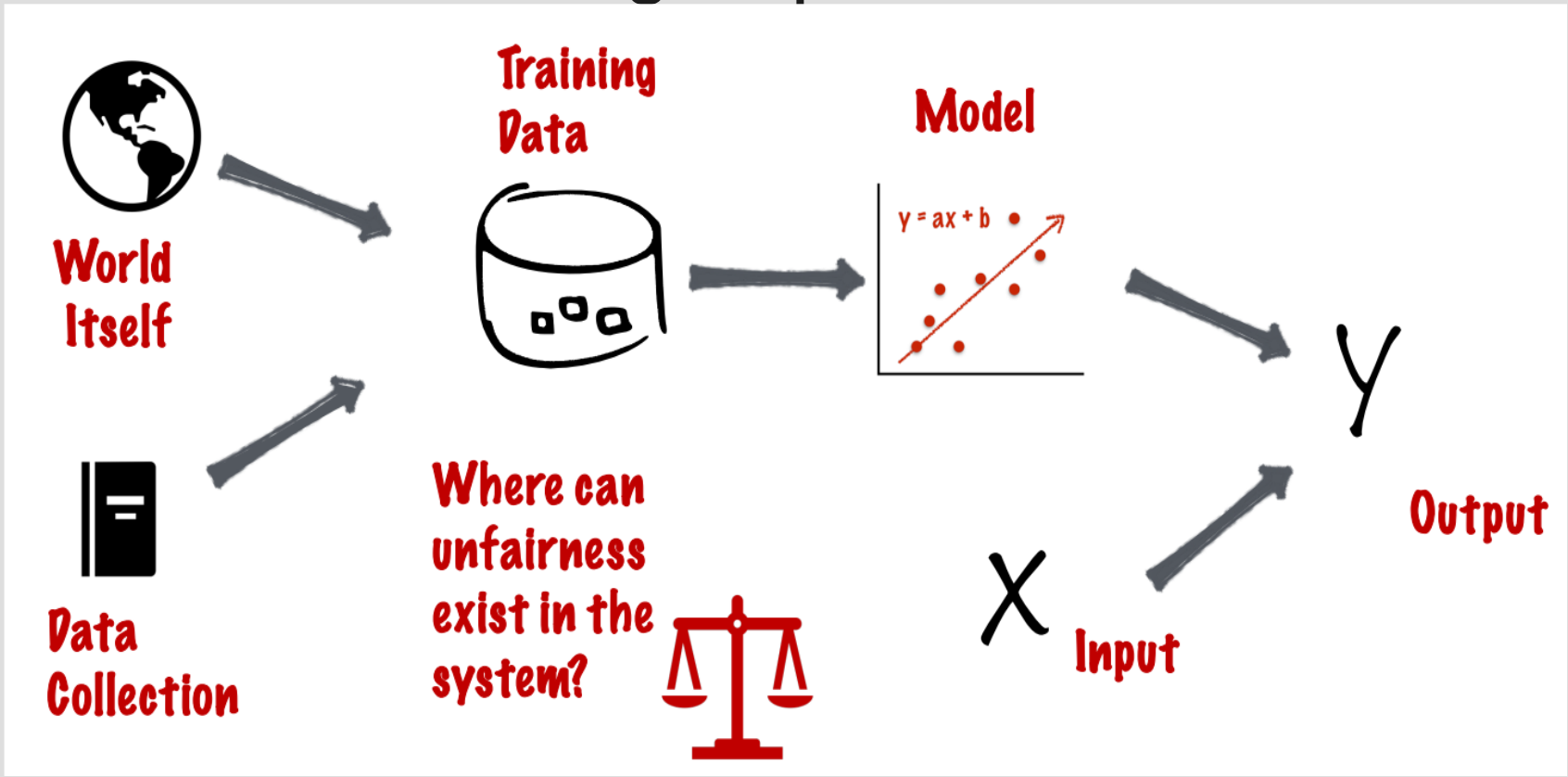
- Demographic Parity: proportion of people who get good outcome/bad outcome should be equal across all groups
- Equal False Positive/False Negative Rates (all confusion matrix scores)
- Equalized Odds: The protected attribute and the prediction are conditionally independent given the ground truth: i.e., the rates of loan application acceptances should be the same across groups among people who are truly credit-worthy
- Individual fairness constraint: similar people should be treated similarly

Problems with fairness constraints

- They don’t always lead to the fair outcomes you think they should either!

- Refer to Measure and Mismeasure of Fairness (Corbett-Davies and Goel) and Delayed Impact of Fair ML (Liu and Hardt)

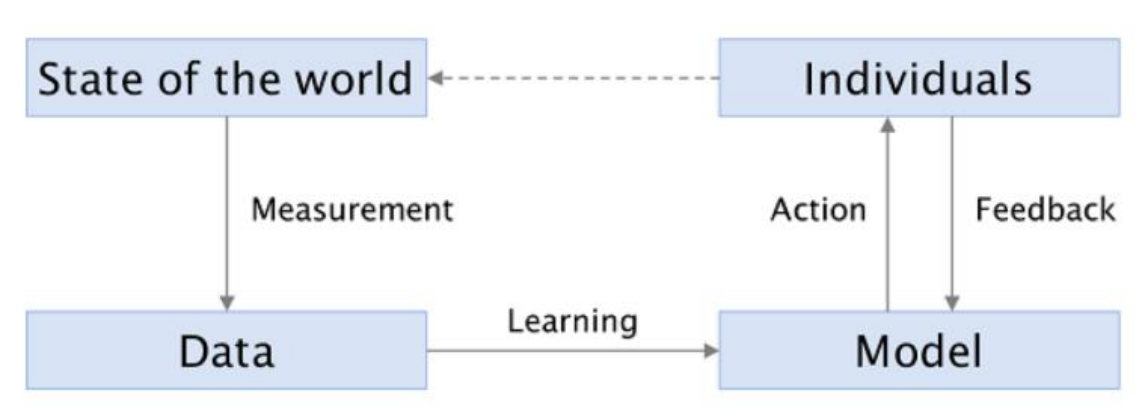
Machine Learning Pipeline



Fairness in ML : Goals

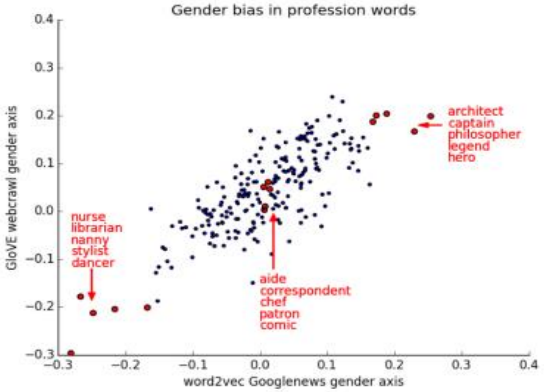
– Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

Stages of ML System



- Measurement: process by which the state of the world reduced to a set of rows, columns, and values in dataset.
- Learning: turns dataset into model
- Action: based on model's prediction (classification, regression, info retrieval), corresponding action
- Feedback: user responses can update model (e.g., clicks)

Unfairness in Machine Learning



Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Joy Buolawmini

The Walrus, 2018

Further Thoughts: Individuals vs Groups

- We can think about fairness in aggregate or individually
- Group fairness: ideas like demographic parity, equalized odds: statistics for all groups should be the same

    --But this doesn't solve all problems

        --What about intersectionality? You could accept the same number of black people and white people to college, but accept no black women

        --Increase disparities within a subgroup: e.g. make it easier for wealthy or otherwise privileged black people to get into college, but make it just as hard or harder for low income students of color

- Individual fairness: similar people should be treated similarly

    --What does it mean for two people to be similar?

Further Thoughts

- ML systems evolve the system that they are deployed in, but ML algorithms do not take this shift into account

- PredPol/ ACLU arguments against its use: sending policemen to already overpoliced areas could further perpetuate the cycle of disproportionate incarceration in America

- But similarly, careless “fair” algorithms could lead to their own problems

  --Consider a ”fair” lending algorithm that lent to the same number of people from groups A and B, where B is disadvantaged. If those in group B are not actually qualified for a loan and default, you actually hurt that population more, and also prevent their being qualified in future because they defaulted

Further Thoughts: Delayed Impact of Fairness

- Liu and Hardt paper, Delayed Impact of Fair Machine Learning

- How can we make fair algorithms that take into account the way they change the data landscape over time?

- What if instead of applying some blindness constraint, or demographic parity constraint, to an algorithm, we instead directly optimize for improving the lives of the affected group over time?

Summary

- Fairness becomes a very popular topic in ML community in recent years
- Fairness matters because it has impact on everyone’s benefit
- Unfairness in ML systems is mainly due to human bias existing in the training data
- No consensus on “the best” definition of (un-)fairness exists

Further Reading

- Barocas, S., Hardt, M. and Narayanan, A., 2019. Fairness and Machine Learning. fairmlbook. Org

- Junzhe Zhang and Elias Bareinboim. Fairness in decision-making — the causal explanation formula. In Proc. 32nd AAAI, 2018

- Fang, Y., Liu, H., Tao, Z. and Yurochkin, M., 2022, October. Fairness of Machine Learning in Search Engines. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 5132-5135).