

Fairness in Machine Learning

Zachary Lipton & Henry Chai

10701 — November 27th

Candidate Screening with Multiple Tests

- In practice, signal may not be fixed
- Employers can conduct multiple interviews to gain more information
- Candidates stream from infinite pool, each either *skilled* or *unskilled*
- Employer wants to minimize # of interviews also worried about FPR
- Interviews give noisy signal, assumed cond. indep. given skill
- Optimal policy to minimize tests per hire subject to FP constraint:
Reject if $<$ prior; Accept if posterior $>$ threshold determined by FP constraint
- Analysis follows game of ruins, random walk on log posterior odds

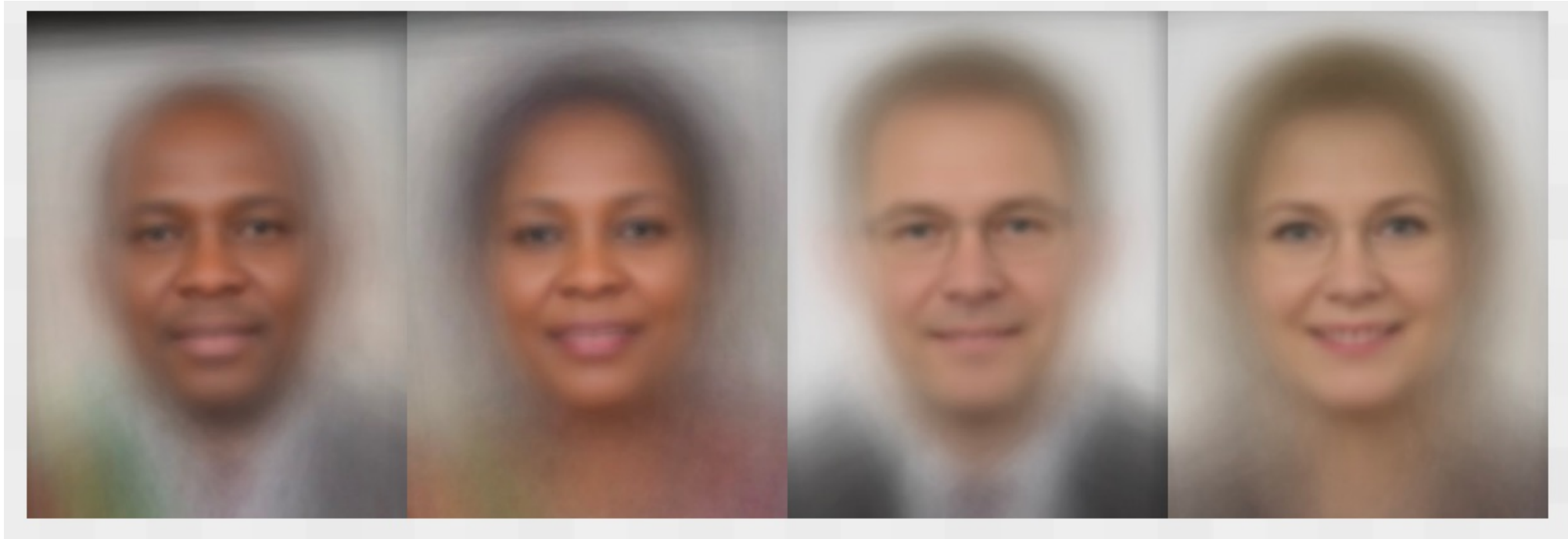


Fair Machine Learning

ProPublica — Machine Bias, 2016



Gender Shades—2018



Bias in word embeddings, 2016

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jameszou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using

Biased allocation of healthcare (2019)

THE VERGE

TECH ▾

REVIEWS ▾

SCIENCE ▾

CREATORS ▾

ENTERTAINMENT ▾

VIDEO

MORE ▾

POLICY

REPORT

SCIENCE

A health care algorithm affecting millions is biased against black patients

A startling example of algorithmic bias

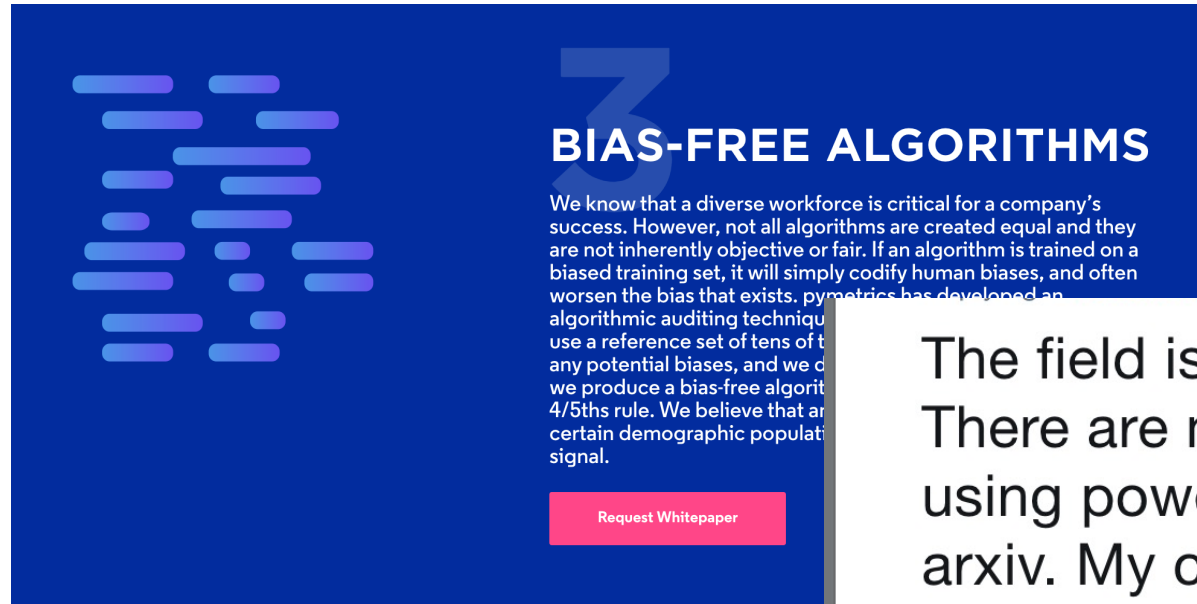
By [Colin Lecher](#) | [@colinlecher](#) | Oct 24, 2019, 2:00pm EDT

“The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients.”

Pernicious Pattern

1. Take a problem ill-described as statistical prediction.
2. Fashion a surrogate prediction problem anyway.
3. Define metrics of success, e.g. accuracy, assuming prediction as task.
4. Trouble arises due to insufficiency of problem description.
5. Work to “solve” the problem while working entirely within the paradigm whose insufficiencies are themselves the root cause.
6. Mislead the public by purporting to have addressed the problem, often by redefining the objective.

Some examples:



3 BIAS-FREE ALGORITHMS

We know that a diverse workforce is critical for a company's success. However, not all algorithms are created equal and they are not inherently objective or fair. If an algorithm is trained on a biased training set, it will simply codify human biases, and often worsen the bias that exists. pymetrics has developed an algorithmic auditing technique that uses a reference set of tens of thousands of people to identify any potential biases, and we do not use any of the data we produce a bias-free algorithm. We believe that a 4/5ths rule. We believe that a certain demographic population signal.

[Request Whitepaper](#)

The field is a bit more sophisticated than this. There are many excellent papers on bias, eg using powerful tools for causal reasoning on arxiv. My colleagues are making good progress and not giving up.

3:46 PM - 30 May 2019

The foundations of *algorithmic bias*

Even if we truly were addressing a prediction problem, things go wrong:

- Some groups under-represented, benefits of automation unequal.
- The training labels themselves may be noisy or *biased*.
- Models often optimized for wrong task altogether (choice of surrogate task may have disparate effects).
- Task may be *easier* for one group.

Complications

- All of our features are correlated.
- And many subject to measurement error.

Anti-discrimination law



President Lyndon B. Johnson shakes hands with Martin Luther King after signing the Civil Rights Act of 1964

Disparate treatment

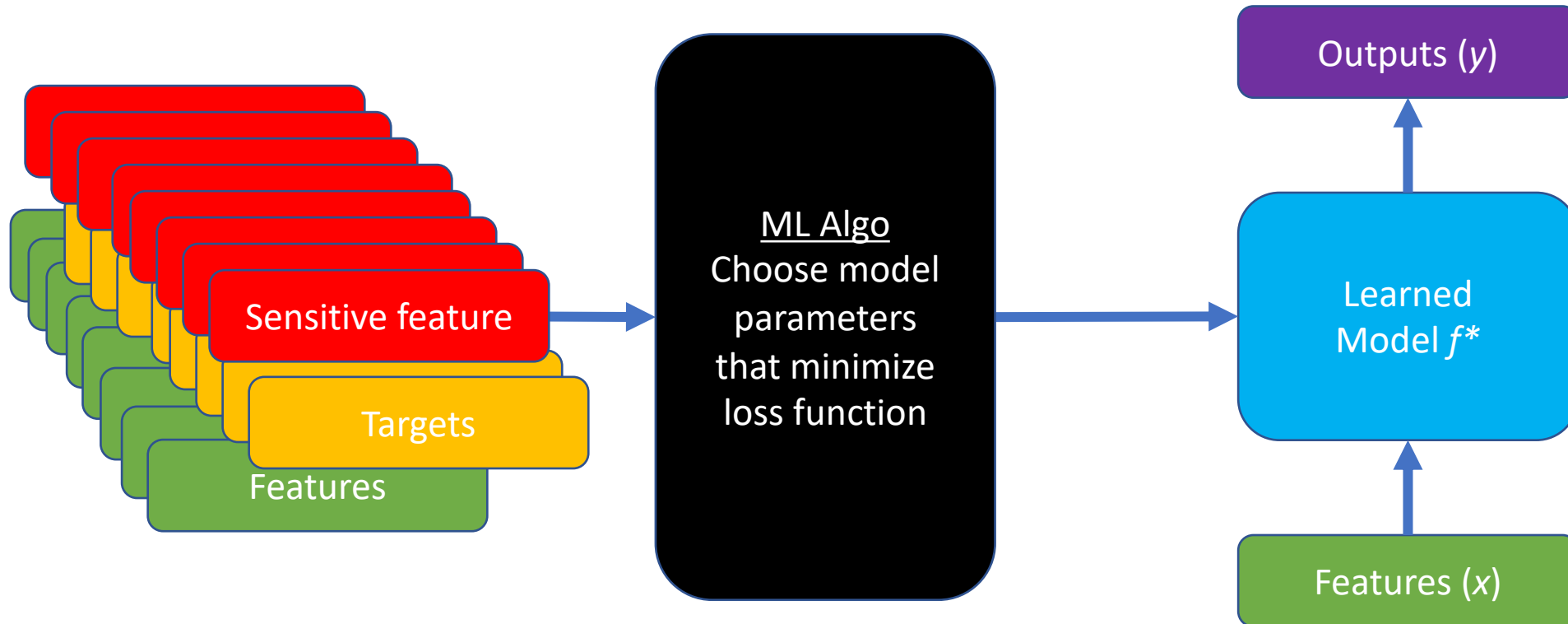
- Addresses intentional discrimination
- Includes decisions explicitly based on a *protected characteristic*
- Also intentional discrimination via proxy variables



Disparate impact

- Facially neutral practices that might nevertheless have an *“unjustified adverse impact on members of a protected class”*
- Complicated doctrine w 3 tests
 1. Plaintiff must demonstrate **statistical disparity** (e.g. 4/5 rule)
 2. Defendant must show that decisions are justified by **‘business necessity’**
 3. Plaintiff must show defendant can achieve goal w **‘alternative practice’**

Fair supervised learning



Make groups equal but how?

- Impact parity
 - Outcome independent of group status $y \perp z$
- Treatment parity
 - The output y depends only on x , not on z
- Representational parity
 - Map x to $r(x)$ such that $r(x) \perp Z$
 - Entails impact parity
- Calibration:
 - Independence of truth and demographic for predicted value — $(T \perp Z | Y)$
- Equalized Odds / “Opportunity” parity
 - Equal false negative and/or false positive rates
- “21 definitions of fairness” (2016)

Impossibility Theorems

- The following 3 conditions cannot (in general) hold simultaneously:
 - Demographic parity ($Y \perp Z$)
 - Separation ($Y \perp Z \mid T$)
 - Calibration ($T \perp Z \mid Y$)
- Characterized by
 - [Chouldechova \(2016\)](#)
 - [Kleinberg, Mullainathan, Raghavan \(2016\)](#).
- Trade-offs among parities unavoidable.

Proliferation of Fair ML metrics and methods

TL;DS - 21 fairness definition and their politics by Arvind Narayanan

2019-07-19 | #fairness , #tl;dr

These are the
give

ir politics
in 2018 by
Princeton

Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions

Shira Mitchell
Civis Analytics
sam942@mail.harvard.edu

Solon Barocas
Microsoft Research and Cornell University
sbarocas@cornell.edu

Kristian Lum
University of Pennsylvania
kl1@seas.upenn.edu

Eric Potash
University of Chicago
epotash@uchicago.edu

Alexander D'Amour
Google Research
alexdamour@google.com

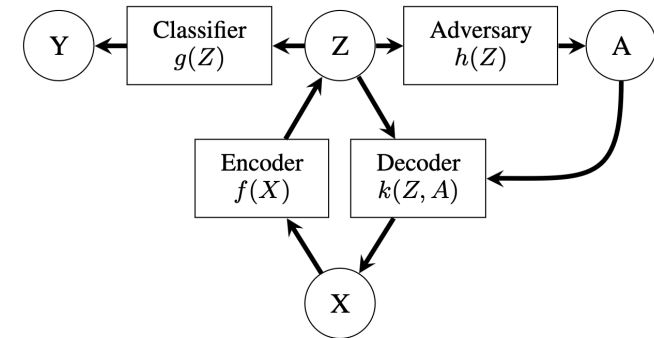
April 28, 2020

April 2020

to define "fairness" for decisions
of this new field has
loguing and

Pre-processing methods

- Somehow treat the data before training with the hope that the pre-processing will ensure some notion of fairness in the end
- Learn “fair representations”
 - Probabilistic approach (*Zemel et al, ICML 2013*)
 - Adversarial learning version (*Madras et al 2018*)
- Flipping negative labels in disadvantaged group (*Toon & Calders 2009*)
- Post-hoc thresholding
(*Corbett Davies 2018, Menon et al 2018, Lipton et al 2018*)



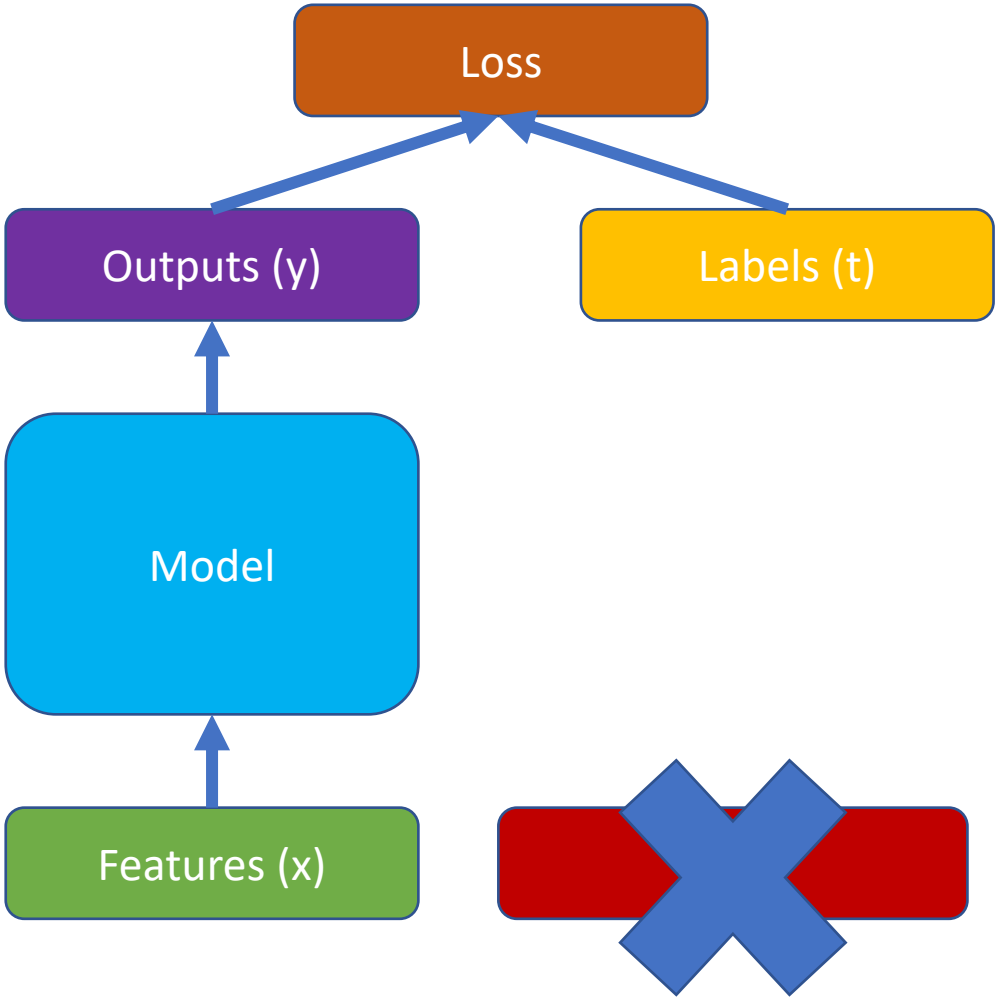
Training with Constraints / Penalties

- Soft penalties / constraints to move decision boundary, enforce some desired parity on the training data
- Margin-based fairness constraints (Zafar et al 2017)
- Mapping parity constraints to “orthogonality” constraint on kernel-based methods (Donini et al 2018)

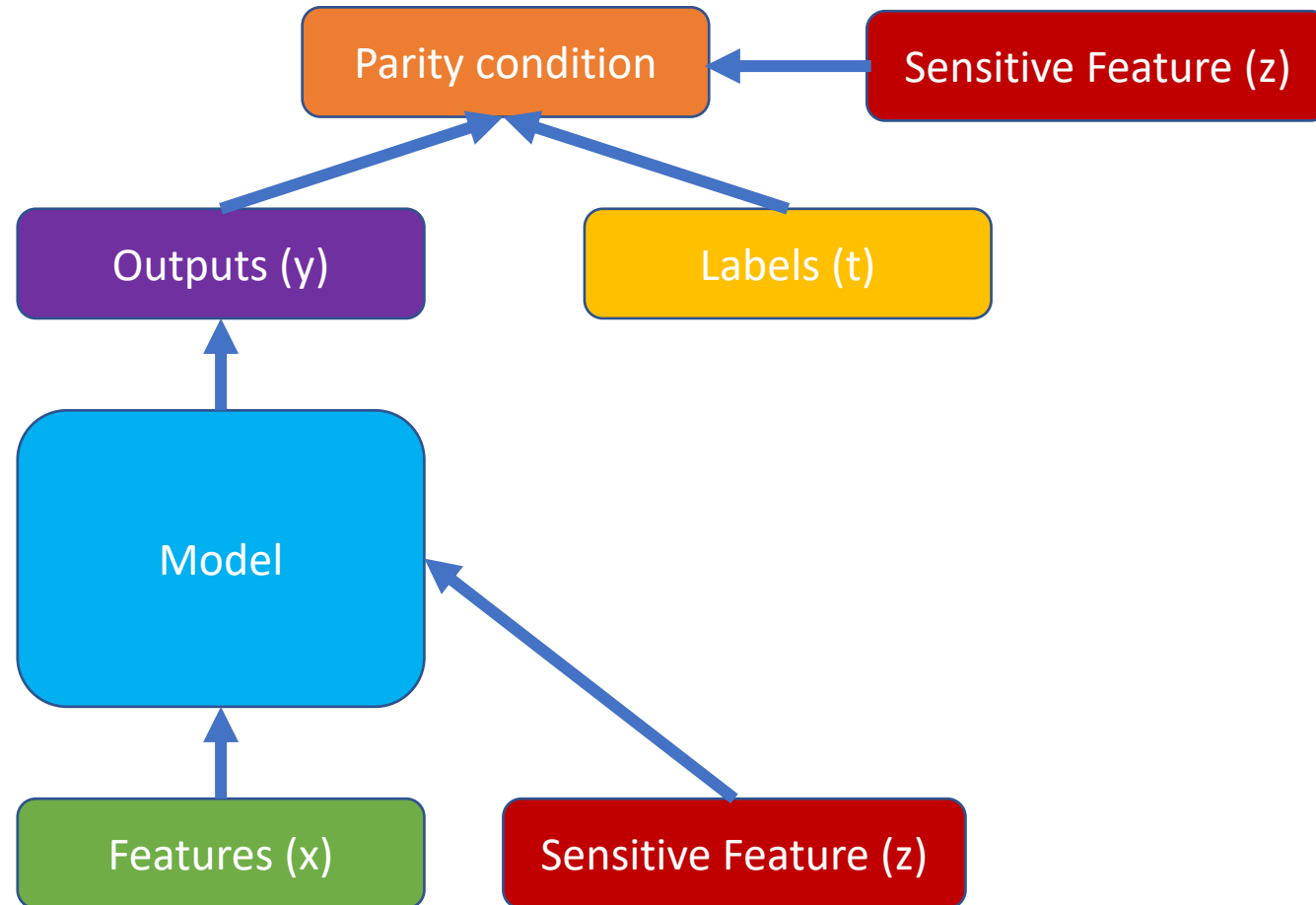
Paper Recipe

- Choose machine learning task T , on data distribution P
- Pick a parity metric M (chosen for mathematical convenience)
- Modify the objective for T or add a constraint to mandate that some functional $f_m(P(\cdot | g=0)) = f_m(P(\cdot | g=1))$
- Call the paper “Fair T ”
 - FairGAN
 - Fair PCA
 - Fair k-means clustering
 - Fair Hierarchical clustering
 - Fair Deep RL
 - Fair bandits

Treatment parity / blindness



Demographic parity / equal outcomes



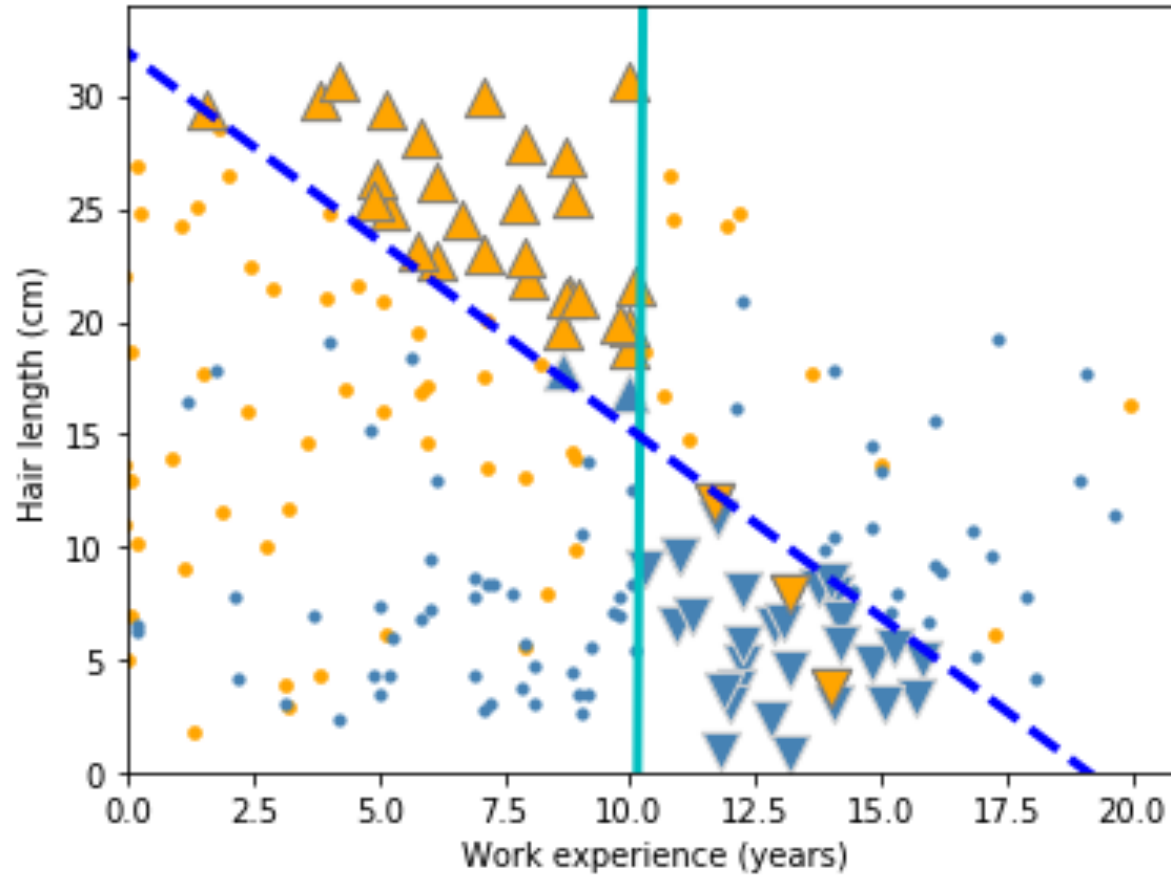
Problems

- If all groups are the same in every way, easy
- Otherwise various parities are mutually irreconcilable
- Statistical parities don't capture legal /philosophical notions
- Do not address whether decisions are justified
- Lacks **even the ingredients** required to determine just action:
 - How did the data come to be / did disparities arise?
 - What are the impacts of decisions?
 - What are responsibilities of the decision-maker?

Findings

1. For reconciling impact disparity and treatment disparity, **treatment disparity is optimal** (theoretical)
2. When \mathbf{x} fully encodes \mathbf{z} , for sufficiently powerful model, **DLP indistinguishable from treatment disparity** (theoretical)
3. When \mathbf{x} partially encodes \mathbf{z} , DLP results in side effects (empirical)
 - A. Re-orders within-group based on otherwise irrelevant characteristics
 - B. Produces potentially bizarre incentive to conform to stereotype

Toy example

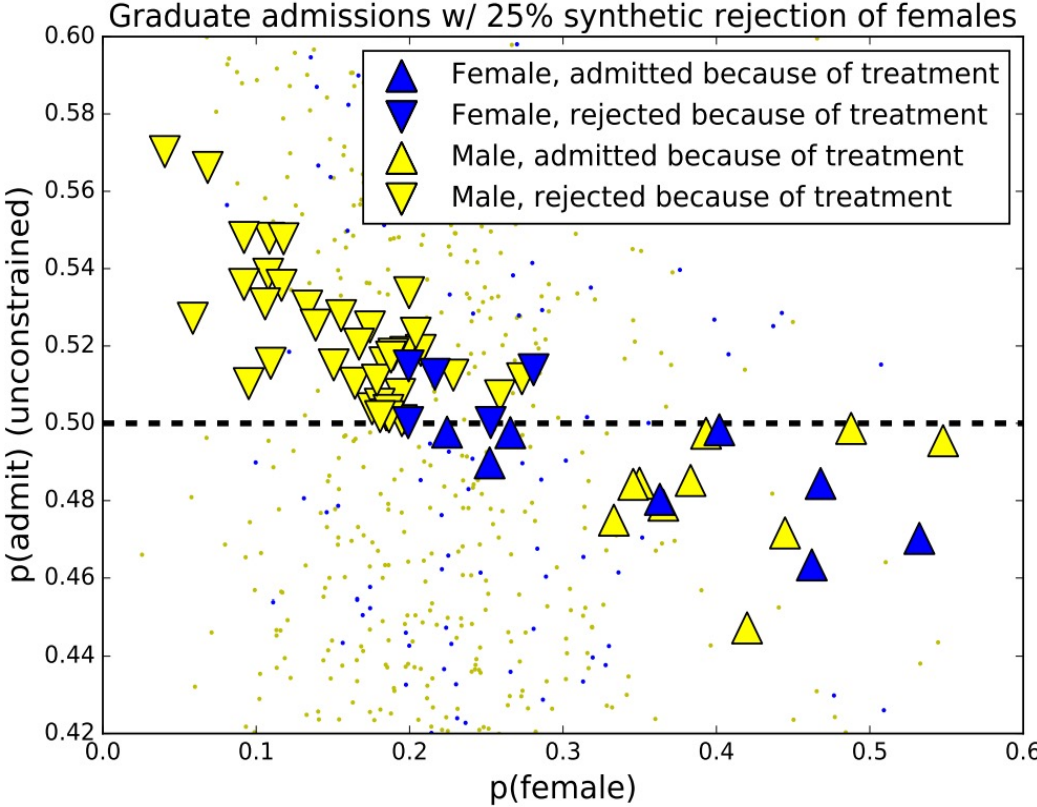
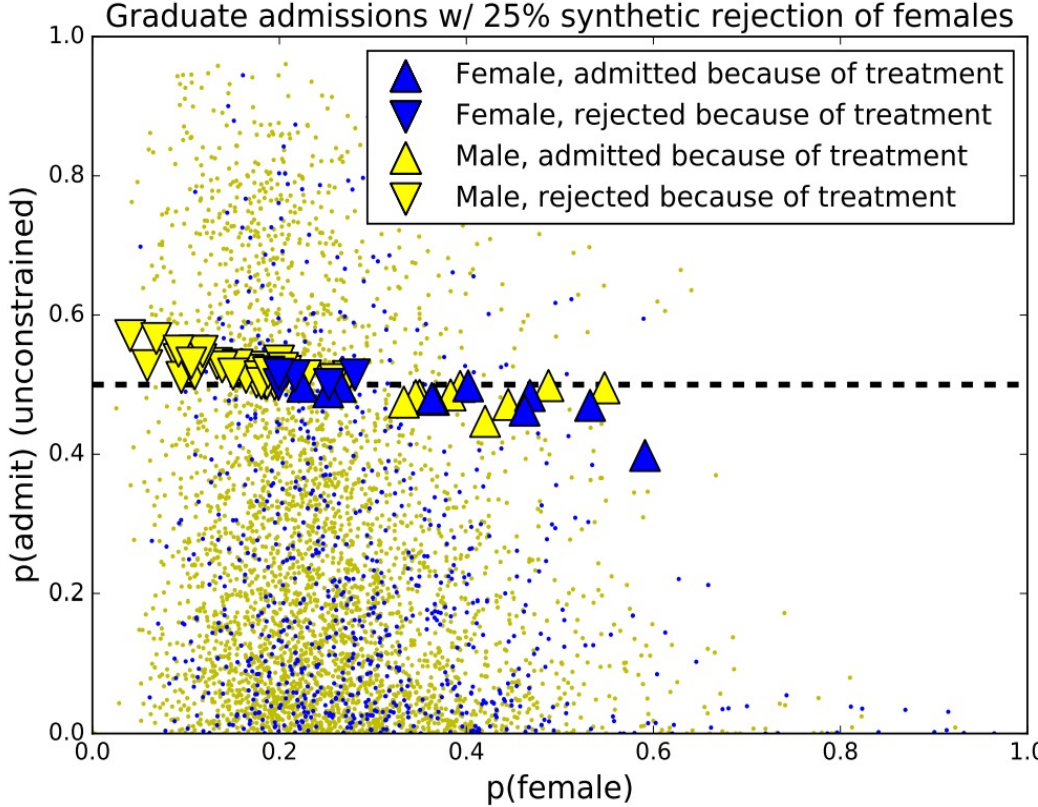


- Acc=0.96; p% rule=26% - Unconstrained
- - - Acc=0.75; p% rule=100% - DLP
- ▲ Woman advantaged by DLP
- ▼ Woman disadvantaged by DLP
- ▲ Man advantaged by DLP
- ▼ Man disadvantaged by DLP

Case study: Gender bias in CS admissions

- **Dataset:** sample of ~9,000 students considered for admission to the MS program of a large US university over an 11-year period
- **Labels:** admissions decisions provided by a faculty admissions committee
- **Attributes:** **Gender** the **protected attribute**. Country of origin, interest area, and GRE, etc. are used as features
- **Synthetic discrimination:** applied to mimic biased training data: of all women who were admitted, we flip 25% of their labels to 0

Effects of DLP in CS admissions



Solutions or Solutionism?

- From the perspective of stakeholders caught in the tension between (i) the potential profit to be gained from deploying machine learning in socially-consequential domains, and (ii) the increased scrutiny of a public concerned with algorithmic harms, these metrics offer an alluring solution: continue to deploy machine learning systems per the status quo, but use some chosen parity metric to claim a certificate of fairness, seemingly inoculating the actor against claims that they have not taken the moral concerns seriously, and weaponizing the half-baked tools produced by academics in the early stages of formalizing fairness as a shield against criticism.

A new perspective on impossibility theorems

- Fair ML clarifies overlooked shortcomings with ideal approaches: In general, if we start from a non-ideal world, no set of actions (by a single agent) can instantaneously achieve the ideal world in every respect. Moreover, matching the ideal in a particular respect, may only be possible at the expense of widening gaps in others.
- This naive form of ideal theorizing is fundamentally underspecified. If matching the ideal in various respects simultaneously is impossible, then we require, in addition to an ideal, a basis for deciding which among competing discrepancies to focus on.

Or... an old perspective on impossibility theorems

Many other problems of applied equity follow a similar pattern. What seems simple at first turns out to be riddled with puzzles and contradictions. Inevitably, we must turn to logical analysis to sort them out. The study of equity turns out, therefore, to have close ties with the axiomatic method in mathematics. From simple and intuitively plausible propositions about the meaning of equity, one draws general and sometimes surprising conclusions about the form that an equitable rule must take.

The axiomatic method has two weaknesses however. The first is that, while each axiom seems reasonable by itself, when piled on top of one another they almost inevitably lead to “impossibility” theorems. This confirms the skeptic’s predisposition to believe that the problem had no solution anyway. The proper conclusion, however, is that not all desirable conditions can be satisfied simultaneously. Some choice must be made. A second difficulty with the axiomatic method is that it can easily become disengaged from the problem that it was intended to solve. The invention of axioms and conditions is a fascinating business. The danger is that the exercise can take on a life of its own and lead to results that are mathematically elegant, but that have little or no relation to the realities of the underlying situation. To guard against this tendency I have tried to mix formal definitions and theorems with informal arguments and examples

weaknesses
of axiomatic
method
interpreting
“impossibility”
theorems

—Peyton Young in “Equity” → 1994!

Causal approaches to fairness

- Counterfactual fairness (Kusner, 2017):
 $P(Y | \text{do}(A=a), \mathbf{x}) = P(Y | \text{do}(A=a'), \mathbf{x})$
- The Causal Explanation Formula (Zhang, Bareinboim 2017):
Decompose correlation between Z and Y into *direct causal effect*, *indirect causal effect* (along various paths) and *spurious effect*.
- Relies on causal model—sensitive to misspecification, subjectivity.
- Outsources the key issue to humans: which paths are impermissible?
- Focuses on effect of Z on Y, is this the right notion of causality?
- Diagnoses unfairness but not *whose actions* are unfair
- Doesn't address who has a responsibility to intervene?

Feedback loops

- Some researchers are beginning to study next-step or equilibrium outcomes in a dynamic model that accounts for interaction w society.
- In [*Delayed Impact of Fair ML*](#), Liu et al. show satisfying certain fairness criteria (wrt static view) can lead to greater long-term disparities.
- In [*Social Cost of Strategic Classification*](#) and [*The Disparate Effects of Strategic Manipulation*](#), authors explore how agents react to policies—consider disparate ability to manipulate features.
- [*Runaway Feedback Loops in Predictive Policing*](#) considers interaction of policing decisions with data observed for subsequent retraining

Some nuggets from Lily Hu's take

Statements about the admissibility / not of direct effects are confused any direct effect can be “zoomed” in on to reveal mediating factors

“The Path-Specific Effects methodology is complex: Combining a more robust theory of race and a normative theory of discrimination with path-specific causal inference methods requires one to draw a causal diagram that rolls together sociological causal mechanisms with normative judgments about which causes and effects ought to be considered fair.”

“But to mistake the important question of what it takes for a decision process to be fair for a question about what causal mechanisms generate observed racial inequality, and whether those mechanisms are unfair, is to pass over a broad normative landscape of reasons we care about racial equality for the very narrow frame of ensuring procedural fairness in a causal chain.”

“Whatever health Black people “would have had” in some convoluted counterfactual scenario is frankly irrelevant to the question of whether actually existing inequality is a matter of injustice—let alone what can be done to remedy it.”

Takeaways

- The “responsible AI/ML” problems have in common not just a set of stakeholders but a requirement of special treatment because they are **not**, at their core, prediction problems
- They all involve some commitments about what impacts various interventions in the world have had, or would have
- This requires some commitments about mechanism, and leads (inexorably?) towards causal formal conceptualizations
- Causal language has done a lot to help us to talk about these problems coherently
- However, while these formalisms give us a rich set of mathematical machinery, it’s not clear that we can actually use it

Thanks!

- **Coauthors:**

Alex Chouldechova (CMU), Julian McAuley (UCSD),
Sina Fazelpour (UBC/CMU)

- **Links:**

- *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*
<https://arxiv.org/abs/1711.07076> (NeurIPS 2018)
- *Algorithmic Fairness from a Non-Ideal Perspective*
<http://zacklipton.com/media/papers/fairness-non-ideal-fazelpour-lipton-2020.pdf>
(AIES 2020)
- The Mythos of Model Interpretability
(<https://arxiv.org/abs/1606.03490>) CACM 2018 (& ICML WHI workshop 2016)