Algorithmic Fairness from a Non-Ideal Perspective

Zachary C Lipton
Carnegie Mellon University

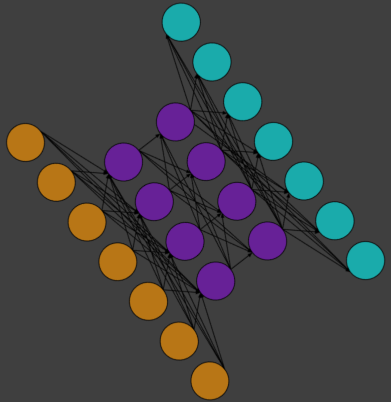
Algorithmic Fairness from a Non-Ideal Perspective https://arxiv.org/abs/2001.09773 (w Sina Fazelpour)
Disparate Learning Processes: https://arxiv.org/abs/1711.07076 (w Chouldechova & McAuley)

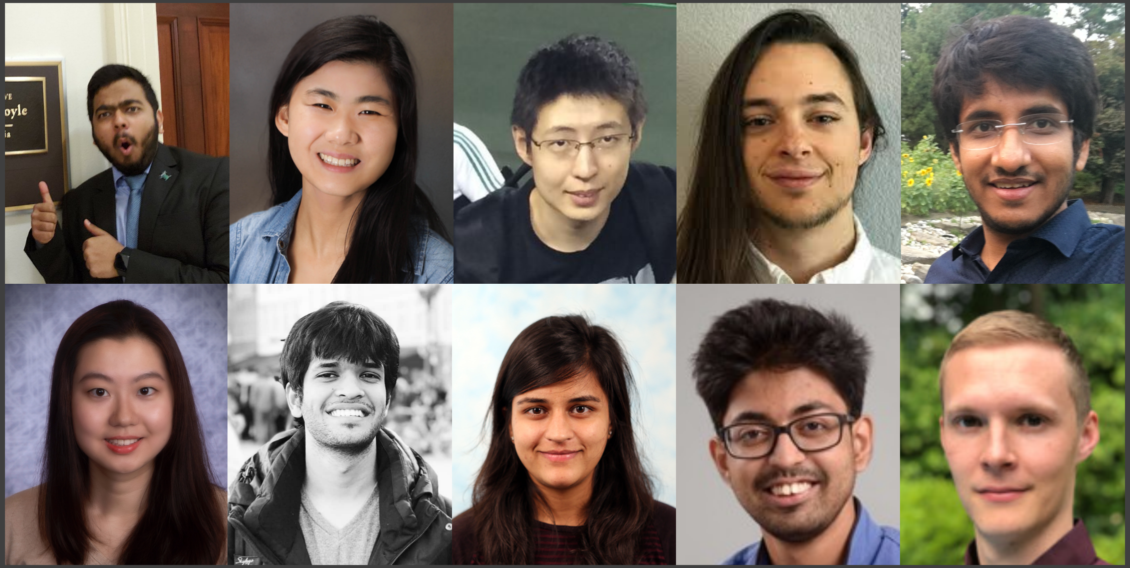
Email: zlipton@cmu.edu
Twitter: @zacharylipton

Approximately Correct Machine Intelligence Lab

- Healthcare
- Robustness
- Causality

- Social Impact & Ethics
- ML / Economics
- NLP





Works and Collaborators

- *Does Mitigating ML’s Impact Disparity Require Treatment Disparity?* https://arxiv.org/abs/1711.07076 (NeurIPS 2018)
- *Algorithmic Fairness from a Non-Ideal Perspective* https://arxiv.org/abs/2001.09773 (AIES 2020)
- The Mythos of Model Interpretability (https://arxiv.org/abs/1606.03490) CACM 2018 (& ICML WHI workshop 2016)

Are these vectors fair?

-8.1, 4.1, 9.6, -3.8, -2.5

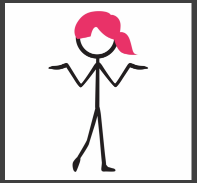
-0.8, 3.4, -7.0, 8.8, -0.8

-0.8, 3.4, -7.0, 8.8, -0.8

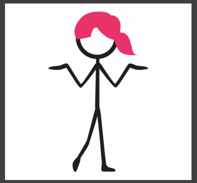
7.6, 1.3, 1.1, -3.2, 8.5

Are these vectors fair?

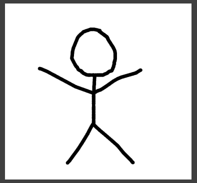
-8.1, 4.1, 9.6, -3.8, -2.5



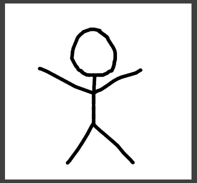
-0.8, 3.4, -7.0, 8.8, -0.8



-0.8, 3.4, -7.0, 8.8, -0.8



7.6, 1.3, 1.1, -3.2, 8.5



Confusion where Technical, Philosophical, Economics, & Legal Terminology Collide

Goals

- Provide conceptual clarity to avoid these category errors
- Put the “fairness” back in “fair machine learning”
- Put the “ethics” back in AI Ethics
- Engage coherently w. political philosophy, economics, & the law
- Examine injustices due to ML & current “fair ML” Critically examine proposed mitigation strategies
- Re-focus attention on **context required** to determine just actions

Institutes of Justinian

The most plausible candidate for a core definition comes from the Institutes of Justinian, a codification of Roman Law from the sixth century AD, where justice is defined as 'the constant and perpetual will to render to each his due'.

— Stanford Encyclopedia of Philosophy

Key points:

- Concerns treatment of individuals
- Arises to resolve conflicts when interests clash
- Justices concerns one's due (an obligation, in contrast to charity)
- Invokes impartiality–two cases relevantly alike should be treated similarly
- Centers on an agent "whose will alters circumstances of its objects"

Conservative vs *Ideal*

- Should justice be viewed “conservative of existing norms & practices” or “demanding reform of these norms and practices”
- Conservative:
  - Respect people’s rights under existing laws, rules & expectations
- Ideal:
  - “Reason to change laws, practices and conventions quite radically, thereby creating new entitlements and expectations”
- The ideal specifies a notion of equality, dismisses claims of justice that do not arise from / accord with the principle.

Corrective vs. Distributive

- Distributive: justice is a principle for allocating good to individuals
  - **Multilateral,** assumes a **distributing agent**
- Corrective: “remedial principle that applies when one person interferes with another’s legitimate holdings”
  - **Bilateral,** concerns relationship between wrong-doer and the wronged
- Idea: theft of a rich person’s property ought to be remediated via corrective justice, but is not demanded by distributive justice
- Philosophers and lawyers disagree about standard of responsibility to mandate corrective justice

Procedural vs. Substantive

- Distinction between the virtue of the method by which benefits and burdens are allocated vs. the final allocation itself.
- Coin tosses may yield equal allocations but be procedurally unjust.
- Some (e.g. Nozick) suggest final distribution is irrelevant, only “sequence of prior events that created it” matters
- Some suggest justice of a procedure is **determined** by its outcomes

Comparative vs. Non-comparative

- When does determining justice require looking at what others can claim?
- Comparative harms:
  - E.g., denied a job that was offered to a less qualified candidate
- Non-comparative harms:
  - Rights to free speech, religion, etc.
  - Whether or not these rights are denied others, they are still one’s right.
- May face trade-offs between comparative/non-comparative harms.
- Focusing myopically on one category can blind us to the other.
- Denying **everyone** a good may have a comparative (but not NC) harm

(Fair ML literature typically focuses on comparative justice)

E.g.: Universal Declaration of Human Rights

- A3—Everyone has the right to life, liberty and security of person. (NC)
- A4—No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms. (NC)
- A5—No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment. (NC)
- A7—All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination. (C)
- A9—No one shall be subjected to arbitrary arrest, detention or exile. (NC)
- A16—Men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to found a family. They are entitled to equal rights as to marriage, during marriage and at its dissolution. (NC & C)

The Scope of Justice

- To who or what does justice apply?
- When & among whom do principles of justice take effect?
- “Who can make claims of justice?”
- “Who might have the corresponding obligation to meet them?”
- If comparative principles are being applied, who should be counted as part of the comparison group?
- Which principles are **universal** vs **contextual**?

Ideal and Non-Ideal Theorizing about Justice

- Key distinction in Rawls and subsequent theorizing on justice/fairness:
- The ideal approach:
  - Imagine a perfectly just world.
  - Try to minimize discrepancy between our world and the ideal.
  - Has been used to argue against affirmative action—*ideal world is color-blind*
- The non-ideal approach:
  - *[Non-ideal theorists] ... seek a* ***causal explanation*** *of the problem to determine what can and ought to be done about it, and who should be charged with correcting it. This requires an evaluation of the mechanisms causing the problem, as well as responsibilities of different agents to alter these mechanisms.*

— "The imperative of integration" Elizabeth Anderson 2019

Algorithmic Fairness from a Non-Ideal Perspective (AIES 2020) https://arxiv.org/abs/2001.09773

Economic perspectives

Becker—“The Economics of Discrimination”

- Considers workers belonging to two groups (say whites & blacks)
- Introduces “taste-based discrimination” a model of outright prejudice
- Employer acts **as though there is a cost** associated w. hiring blacks
- However, profit function π_i regards two groups as perfect substitutes
- Each employer’s utility function assigns “disutility” d_i per black worker

$$V_i = \pi_i - n^b_i \cdot d_i$$

- Market equilibrium results in
  1. Induces a sorting of workers ➔ firms hire only blacks or only whites
  2. Different wages for white and black workers
  3. Wages determined by the marginal discriminator

Arrow’s Rebuttal of Becker (1973)

- Argues that taste-based discrimination will fail because discriminating employers will be driven from market by inefficiency
- Discusses situations with actual productivity differences among groups due to discrimination in other spheres of life (e.g., education)
- Argues that consequence of forcing identical wages may be that employers stop employing from minority group
- Suggests **imperfect information** as alternative cause of disparities

“I believe these results are only the barest fragment of what could be found with better and more detailed systems in which there is an interaction between reality and perceptions of it”

“The statistical theory of racism and sexism”

- Introduced by Phelps (1972)
- Models how disparities arise absent disutility, and w. identically dist. skills.
- Requires only signal more difficult to obtain for minority workers.
- Simplified by Aigner & Cain (1977)
  - Worker quality *q* normally distributed, group-conditioned noise levels *u.*
  - Observed test results $y = q + u, u \sim N(0, \sigma_g^2)$

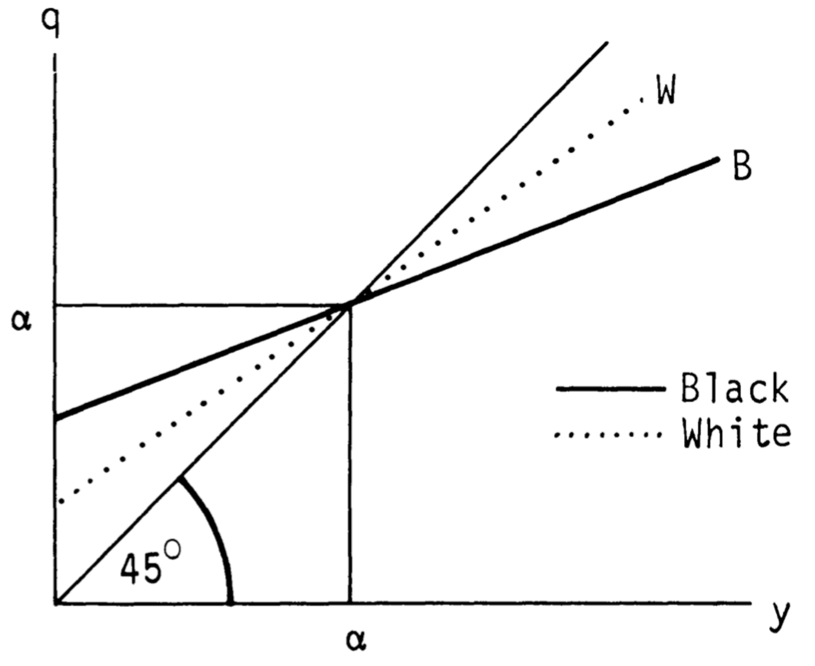


Figure 1B. Predictions of Productivity (q) by Race and Test Score (y), Assuming a Steeper Slope for Whites.

Modeling dynamics of affirmative-action

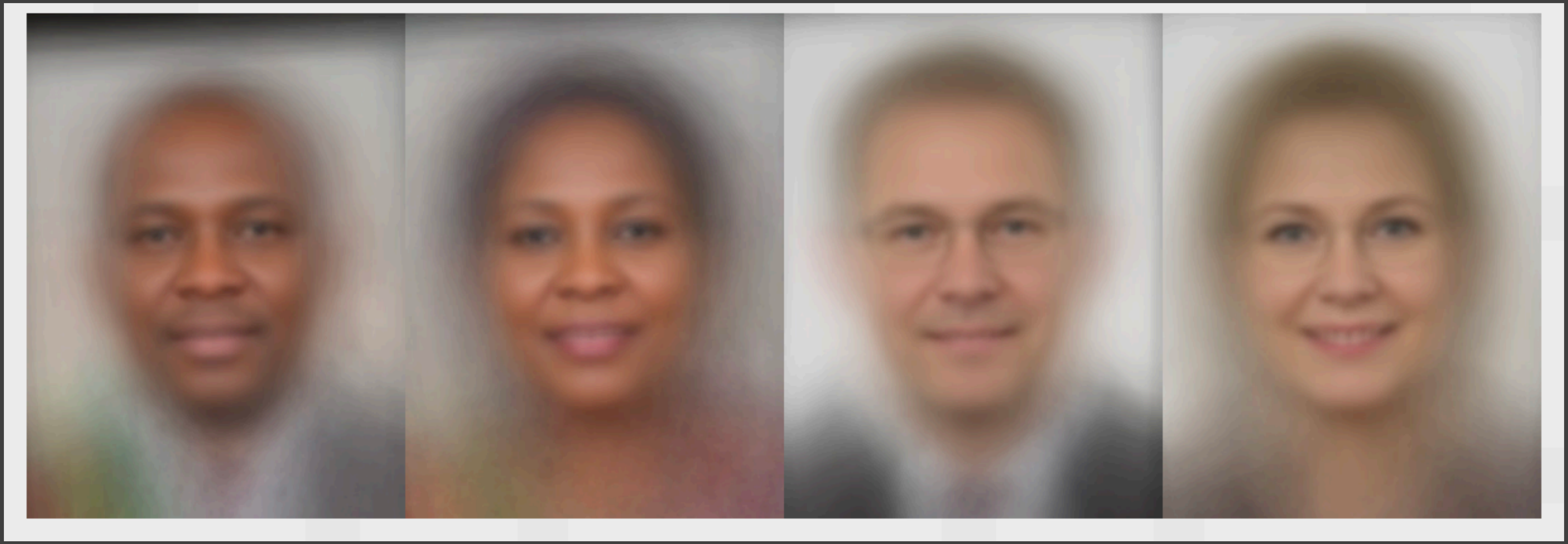
- Several papers and a book (the anatomy of racial inequality) by Glenn Loury investigate discrimination in hiring, examining interplay of policies and employee behavior.
- Coate & Loury (1993) look at long term effects of affirmative action.
- Consider interplay of interventions, investment in education.
- One key insight: even when groups are equal ex ante, equilibrium outcomes following some interventions can **appear** to confirm negative stereotypes.

Will Affirmative-Action Policies Eliminate Negative Stereotypes?
— Coate & Loury (American Economic Review 1993)

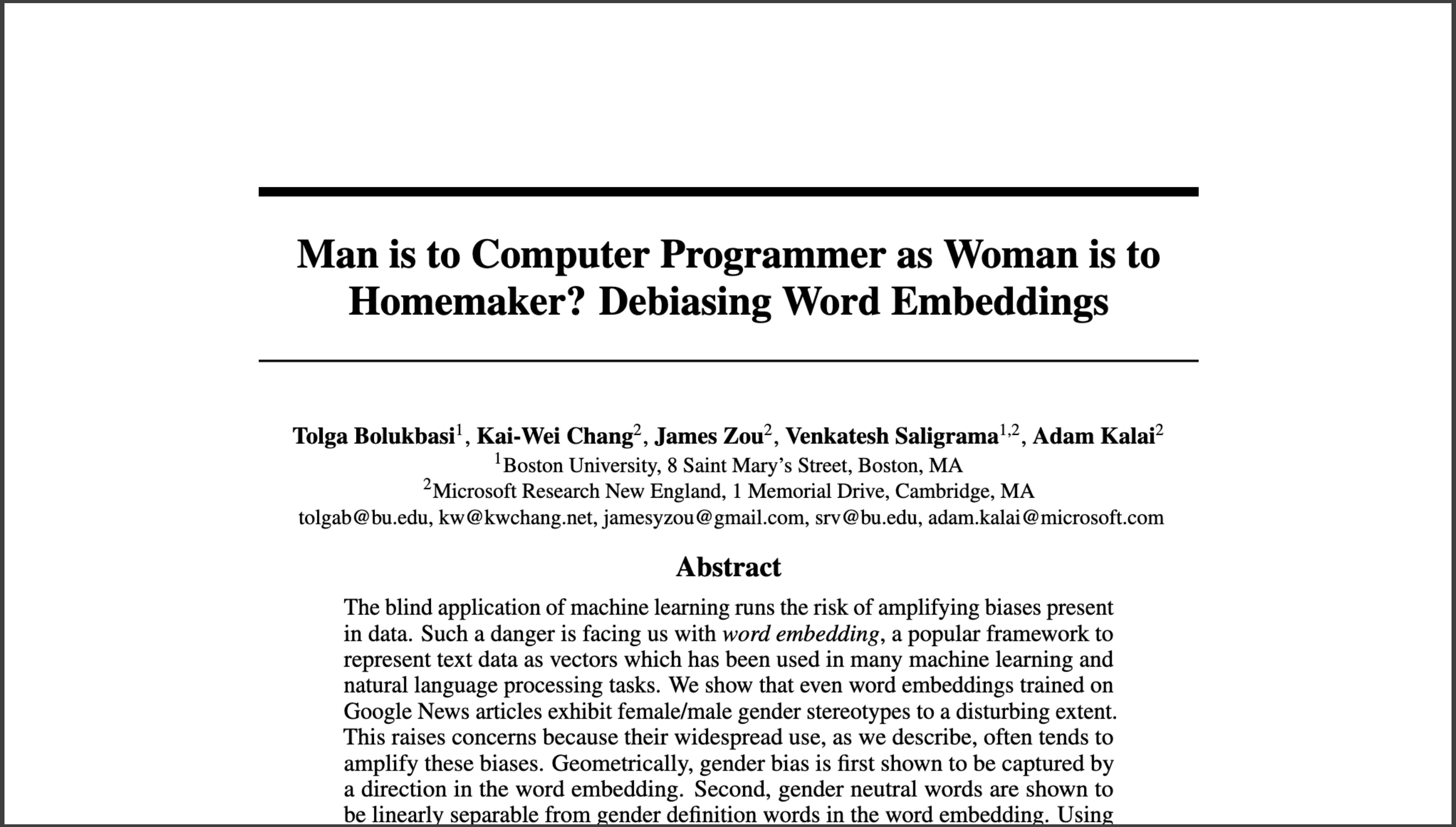
Fair Machine Learning

ProPublica — Machine Bias, 2016

Gender Shades—2018



Bias in word embeddings, 2016



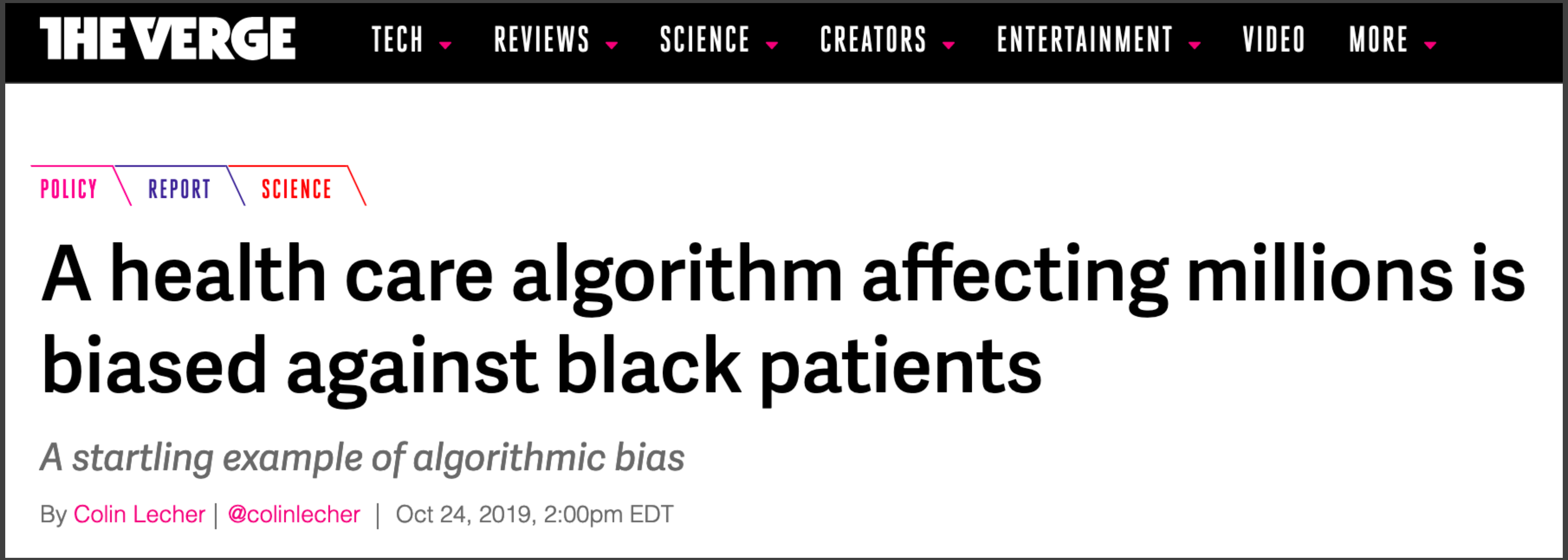
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@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)



POLICY \ REPORT \ SCIENCE

A health care algorithm affecting millions is biased against black patients

A startling example of algorithmic bias

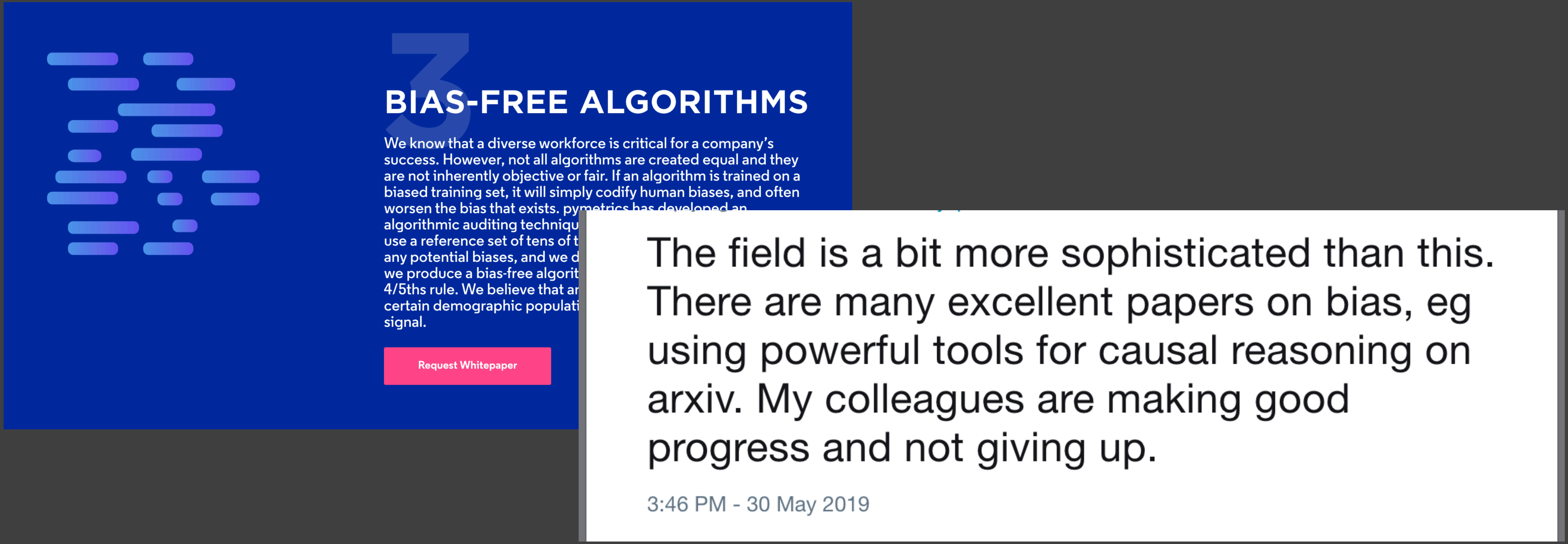
By Colin Lecher | @colinlecher | Oct 24, 2019, 2:00pm EDT

“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:



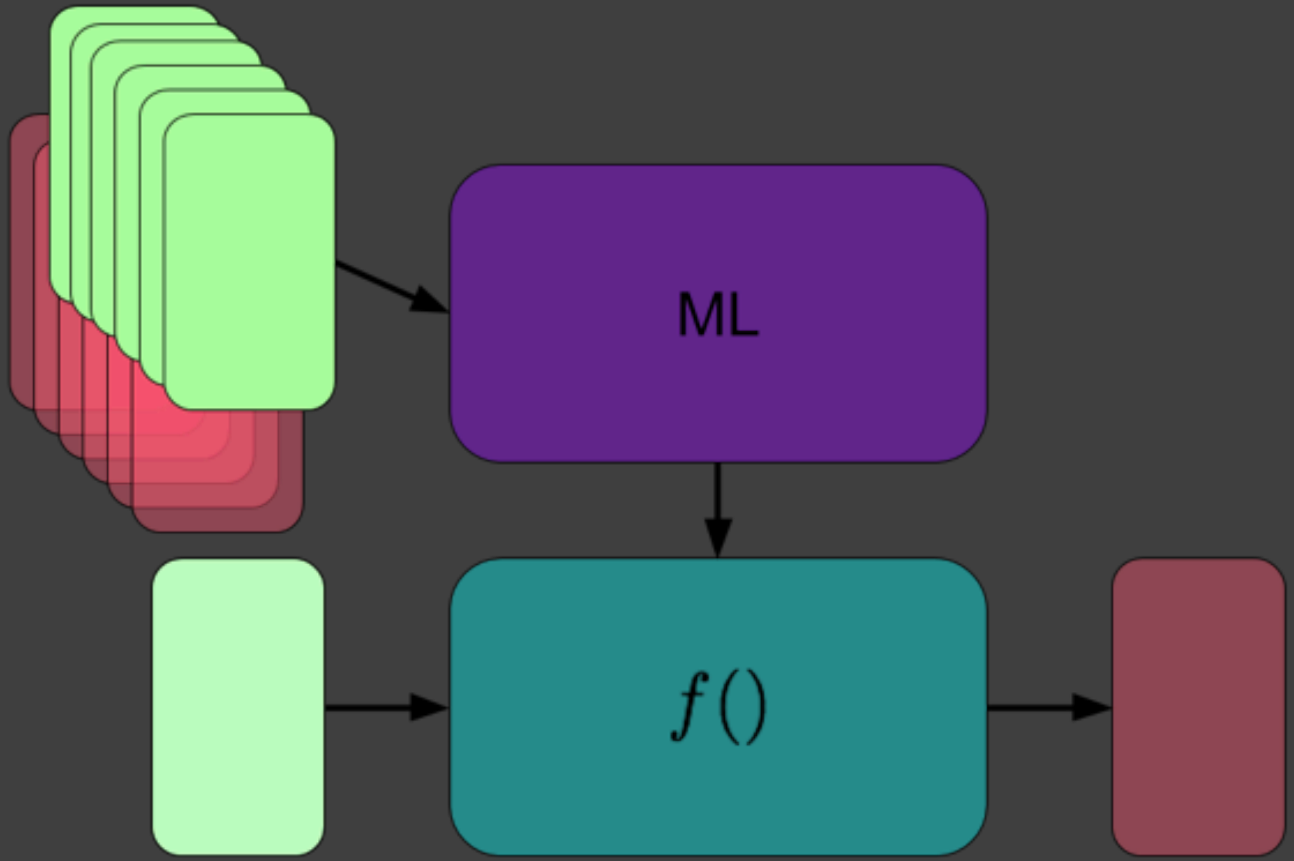
BIAS-FREE ALGORITHMS

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 techniqu use a reference set of tens of t any potential biases, and we d we produce a bias-free algorit 4/5ths rule. We believe that ar certain demographic populati signal.

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.

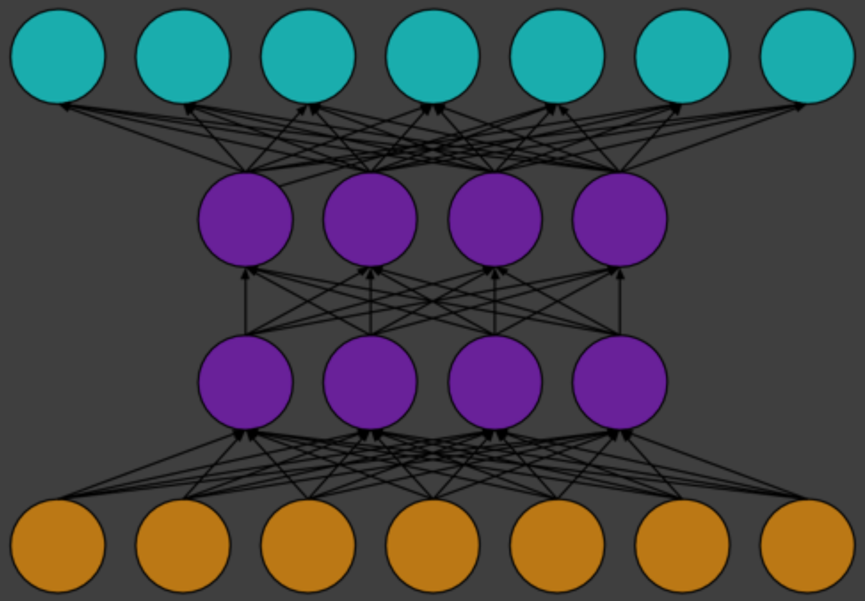
3:46 PM - 30 May 2019

Supervised Learning

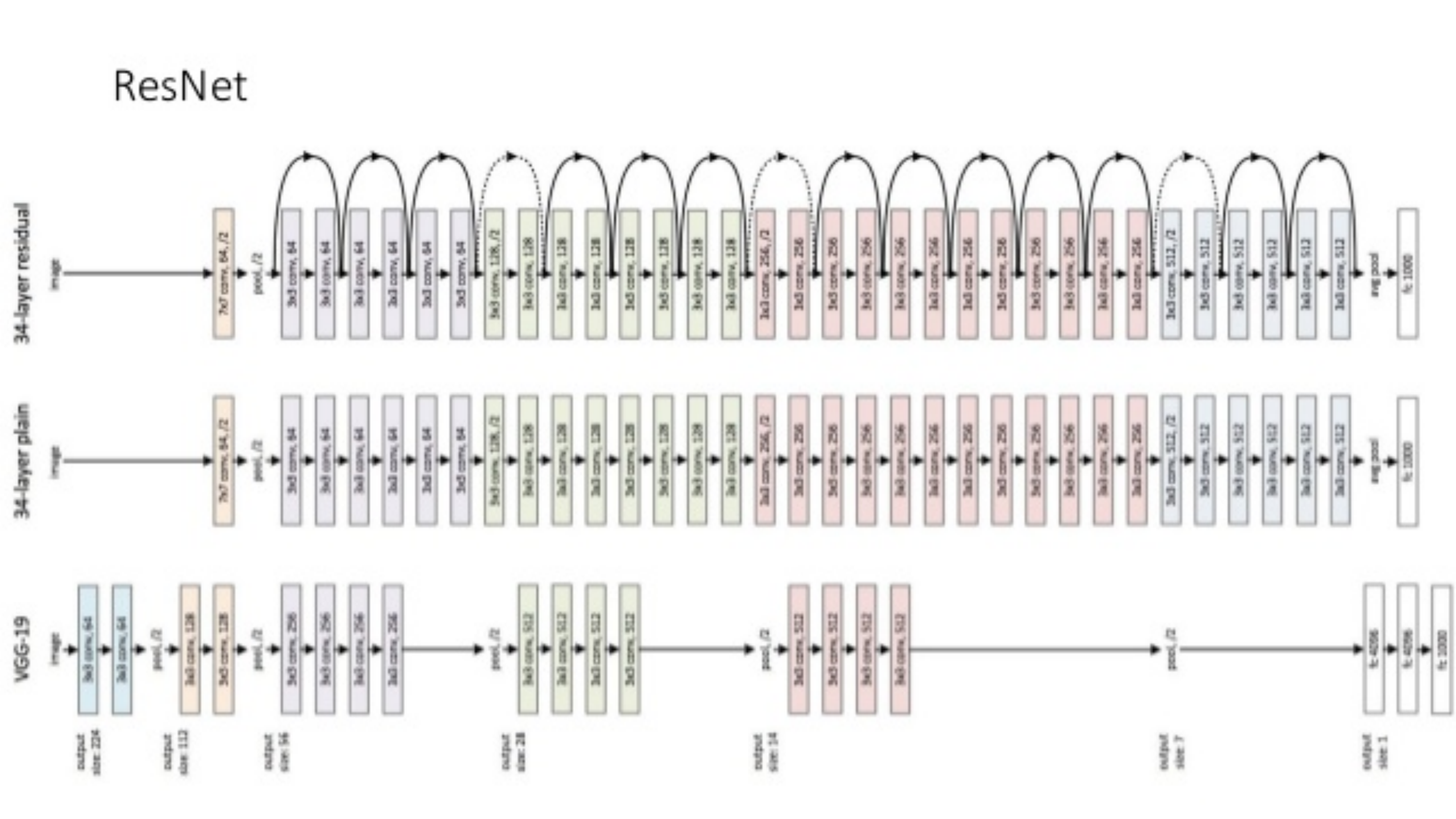


Neural Networks

- Composed of artificial **neurons**
- Connected by weighted edges (*like synapses*)
- Each takes **input**, emits **output**
- Can approximate complex functions

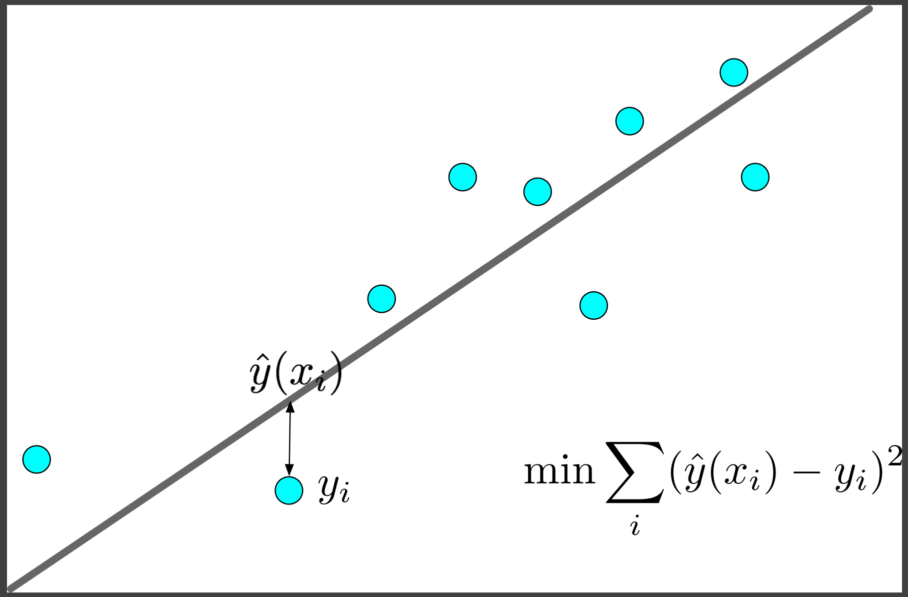


Deep Learning (2005+)



Curve-fitting

- How a pioneer of artificial intelligence became one of its sharpest critics
- “ML is stuck on... learning associations”
- ...we did not expect...so many problems could be solved by pure curve fitting”
- Learns associations, not causal relations
- Sometimes, that’s enough



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

STRINGER/AFP/GETTY IMAGES

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

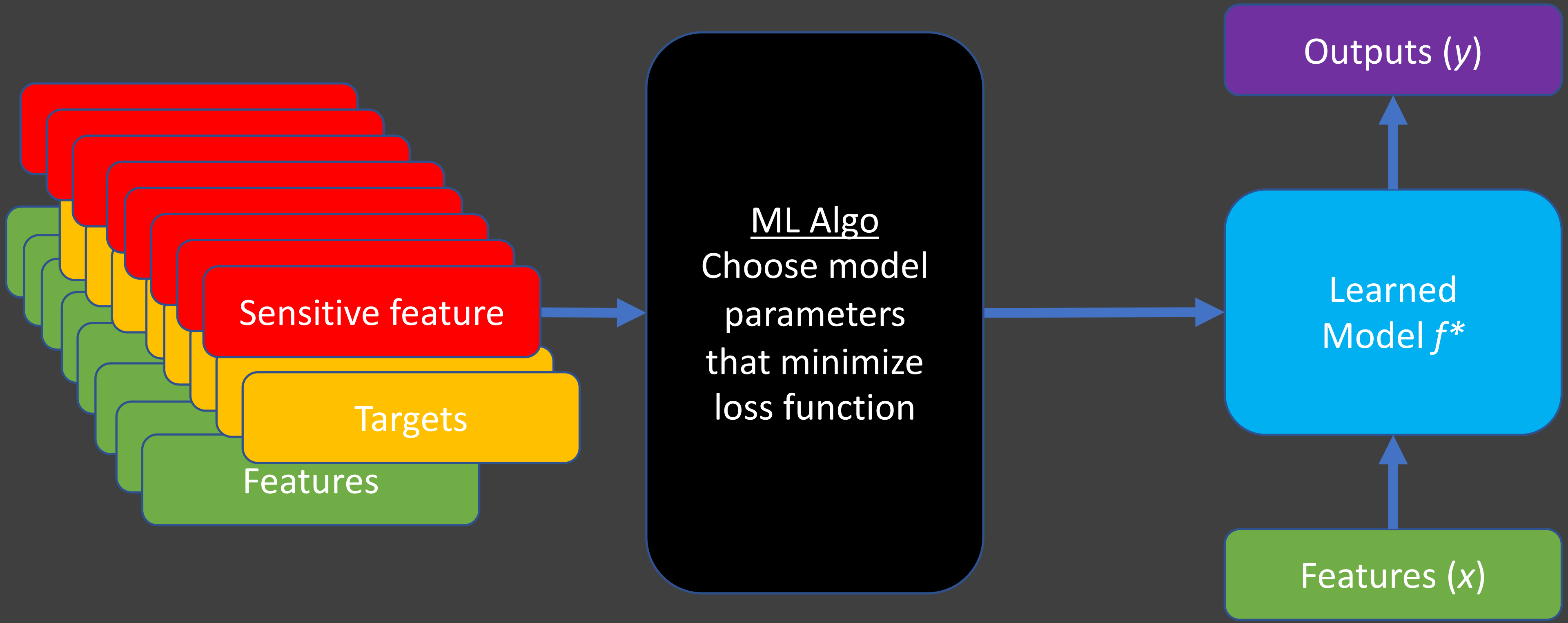
Disparate treatment

- 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’**

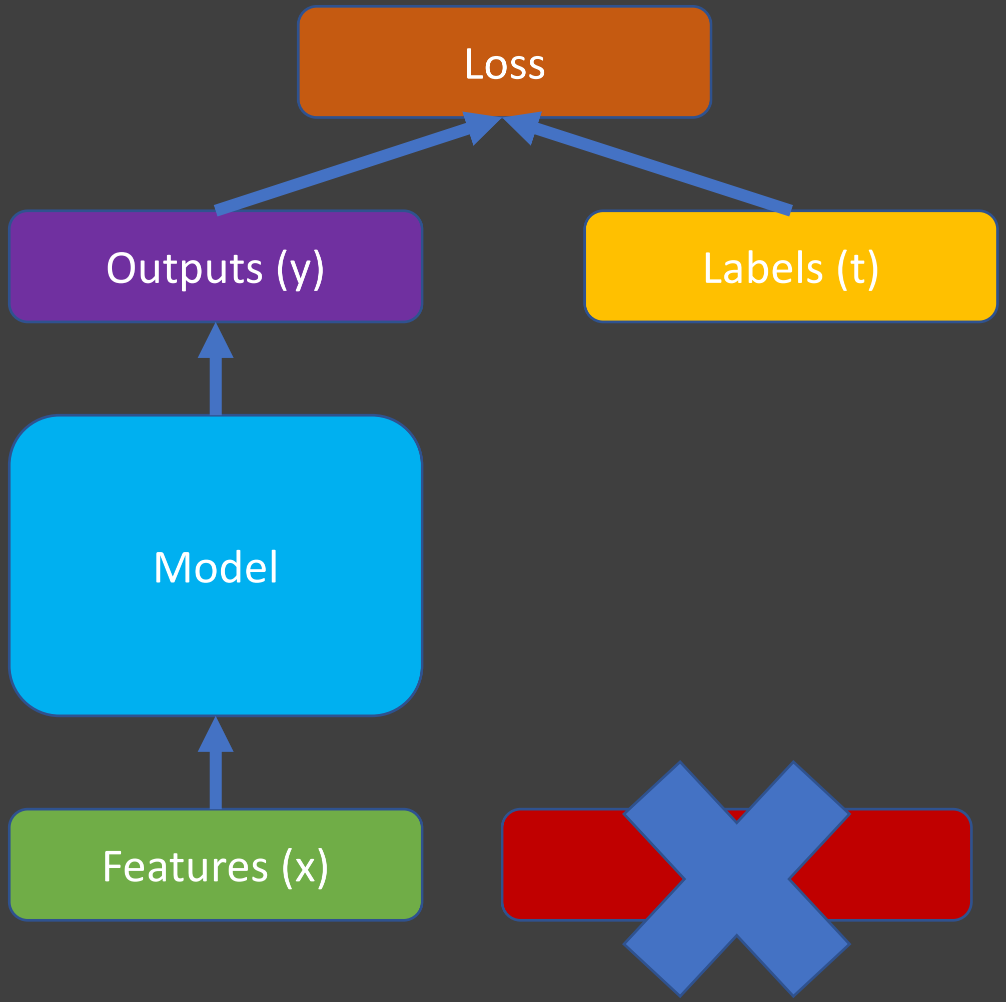
Fair supervised learning



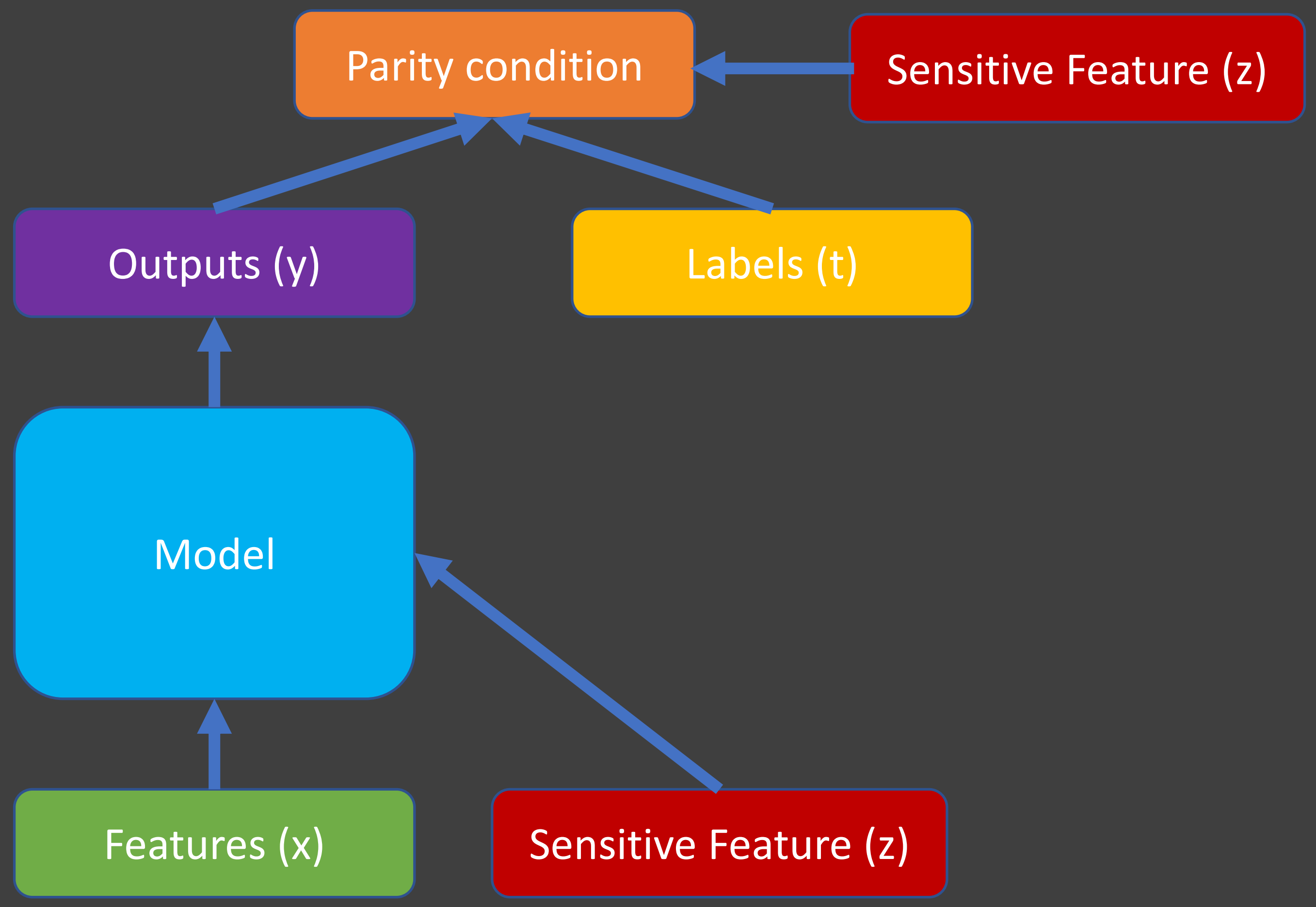
Make groups equal but how?

- 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
- Equalized Odds / “Opportunity” parity
    - Equal false false negative and/or false positive rates

Treatment parity / blindness



Impact Parity / equal outcomes



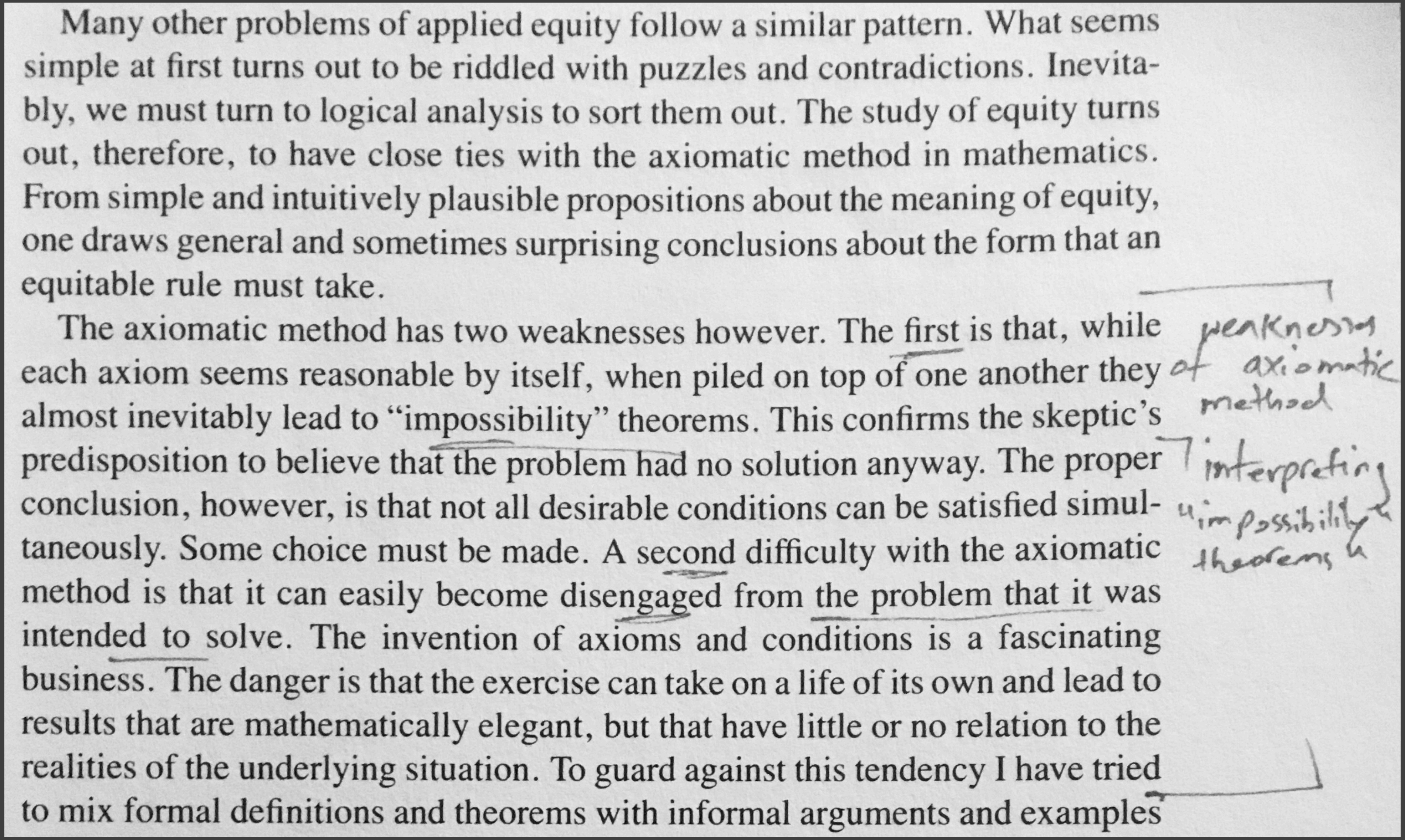
Problems

- 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 came to be / did disparities arise?
  - What are the impacts of decisions?
  - What are responsibilities of the decision-maker?

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.

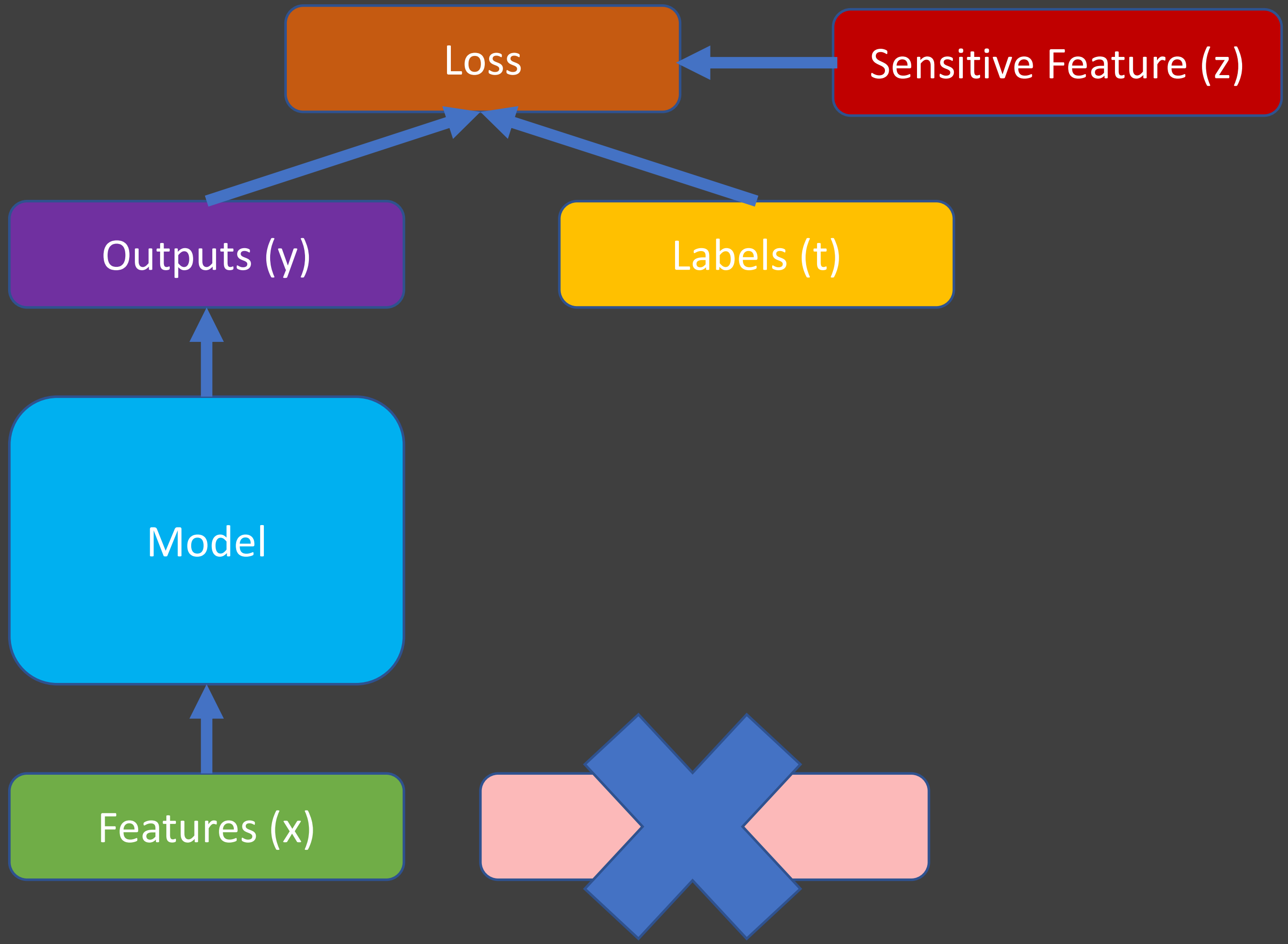
The axiomatic method has two weaknesses however. The first is that, while each axiom seems reasonable by itself, when piled on top of one another they almost inevitably lead to "impossibility" theorems. This confirms the skeptic's predisposition to believe that the problem had no solution anyway. The 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!

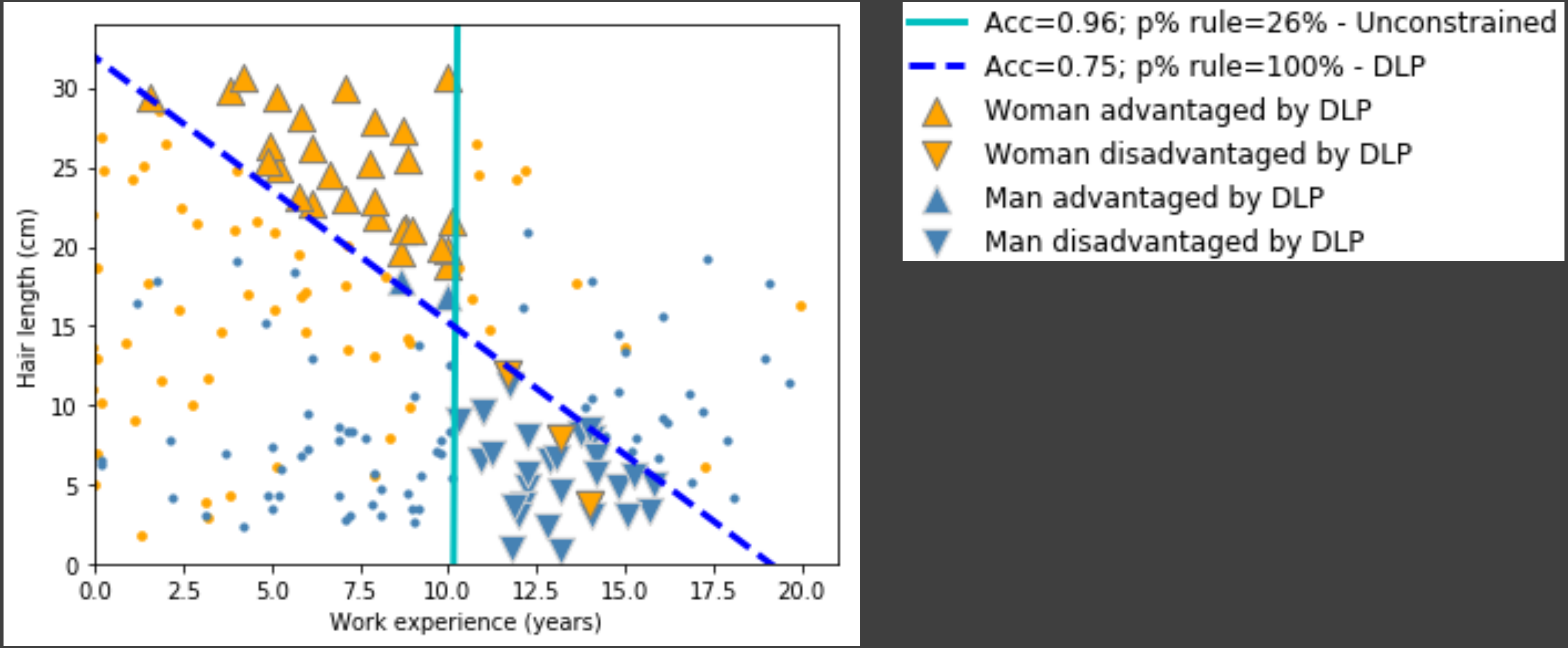
Disparate Learning Processes



Findings

1. For maximizing utility subject to constraint on impact disparity, **treatment disparity is optimal** (theoretical)
2. When ***x*** fully encodes ***z***, for sufficiently powerful model, **DLP indistinguishable from treatment disparity** (theoretical)
3. When ***x*** partially encodes ***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 sterotype

Toy example

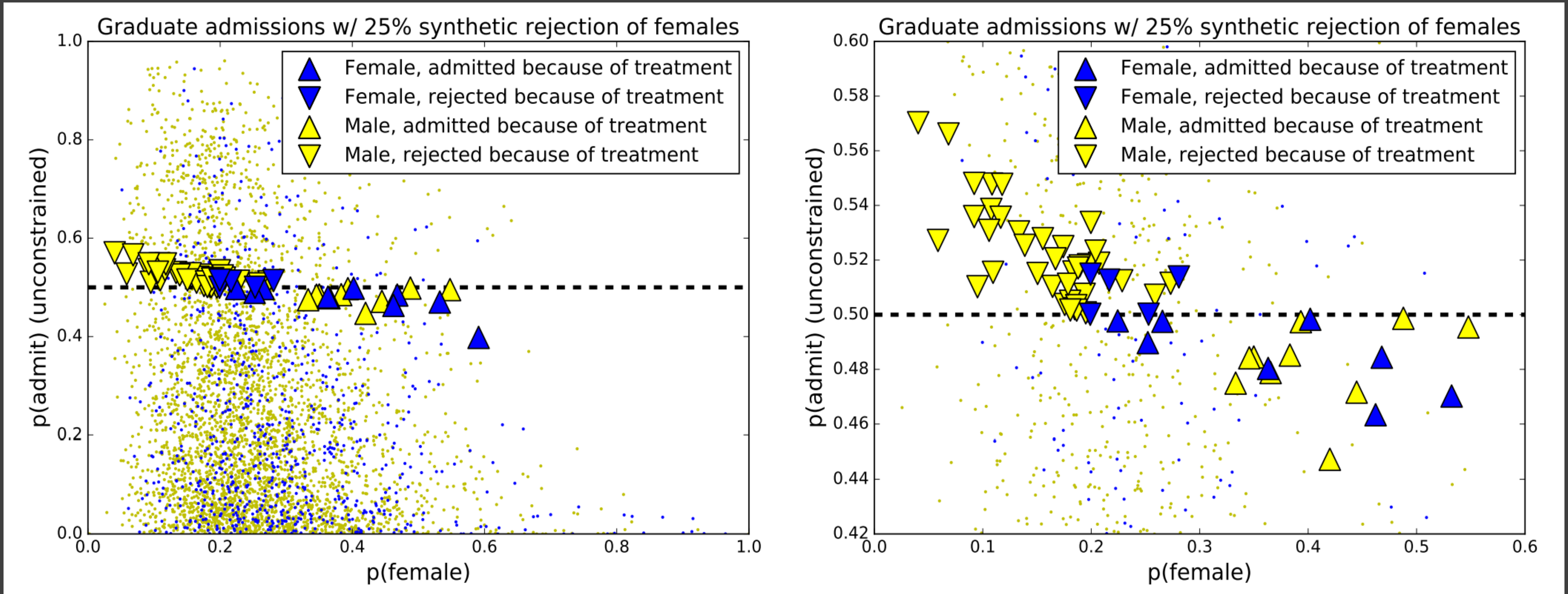


Hair length (cm)
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Work experience (years)
Acc=0.96; p% rule=26% - Unconstrained
Acc=0.75; p% rule=100% - DLP
Woman advantaged by DLP
Woman disadvantaged by DLP
Man advantaged by DLP
Man disadvantaged by DLP

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



Ideal and Non-Ideal Theorizing about Justice

- The ideal approach:
  - Imagine a perfectly just world.
  - Try to minimize discrepancy between our world and the ideal.
  - Has been used to argue against affirmative action—*ideal world is color-blind*

- The non-ideal approach:
  - *[Non-ideal theorists] ... seek a* ***causal explanation*** *of the problem to determine what can and ought to be done about it, and who should be charged with correcting it. This requires an evaluation of the mechanisms causing the problem, as well as responsibilities of different agents to alter these mechanisms.*

    — Elizabeth Anderson 2019 The imperative of integration

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.

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)