Algorithmic Fairness and Anti-Discrimination Law

Alice Xiang
Head of Fairness, Transparency, and Accountability Research
Partnership on AI

About me

- Juris Doctor, Yale Law School
- Master of Science in Economics for Development, Oxford University
- Master of Arts in Statistics, Harvard Graduate School of Arts and Sciences
- Bachelor of Arts in Economics, Harvard College

Currently:

- Head of Fairness, Transparency, and Accountability Research, Partnership on AI

Previously:

- Visiting Scholar, Yau Mathematical Sciences Center at Tsinghua University
- Lawyer, Data Scientist, Econometrics Researcher

PARTNERSHIP ON AI

Who we are

Founded by Google, Microsoft, Facebook, IBM, and Amazon

Consortium of over 100 tech companies, academic institutions, and non-profits

Conduct research at intersection of AI and society

Mission

Bringing diverse voices together across global sectors, disciplines, and demographics so developments in AI advance positive outcomes for people and society.

PARTNERSHIP ON AI

My work

Lead interdisciplinary research team on fairness, transparency, and accountability (FTA) in AI

Conduct and oversee research on:

- Algorithmic Fairness
- Explainable ML
- Criminal justice risk assessment tools
- Diversity and inclusion in field of AI

Overview of Talk

“Reconciling Legal and Technical Approaches to Algorithmic Bias,” to be published in the *Tennessee Law Review*.

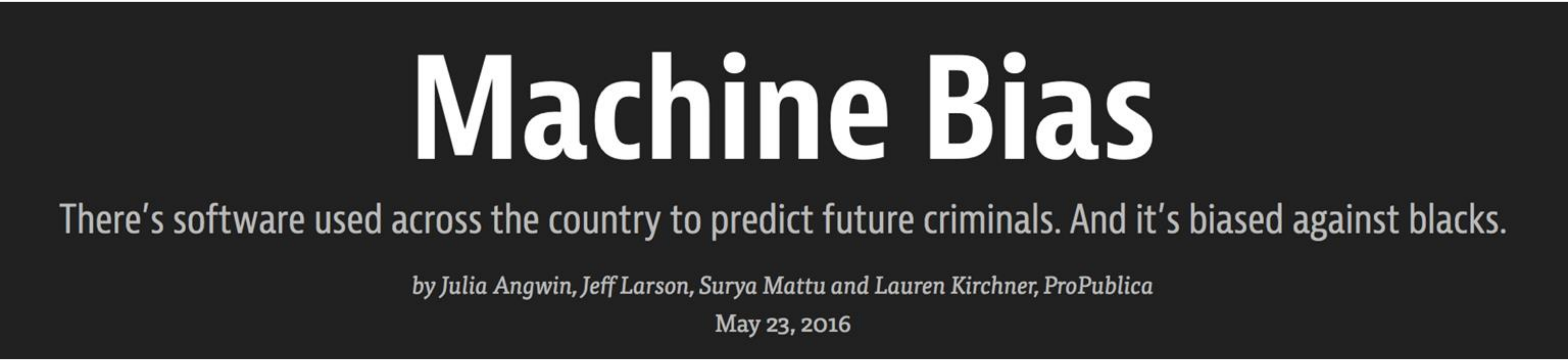
- Key issue: extent to which well-meaning algorithm developers can use protected class variables to address algorithmic bias, in light of existing anti-discrimination law
- Discusses the legal compatibility of technical methods proposed in algorithmic fairness literature
- Today’s talk focuses on setting the stage

Overview of Talk

- Key Tension
  - Technical necessity for using protected class variables to mitigate bias
  - Legal preference for methods that are blind or neutral to protected class attributes
- Agenda
  - What is algorithmic bias?
  - Need for use of protected class variables in mitigating algorithmic bias
  - How principles in anti-discrimination jurisprudence might conflict in algorithmic context
  - Potential role of causal inference in reconciling legal and technical approaches
  - Q&A

Why should we care about algorithmic bias?

In an age of AI,
automated decisions
will increasingly
affect our lives.

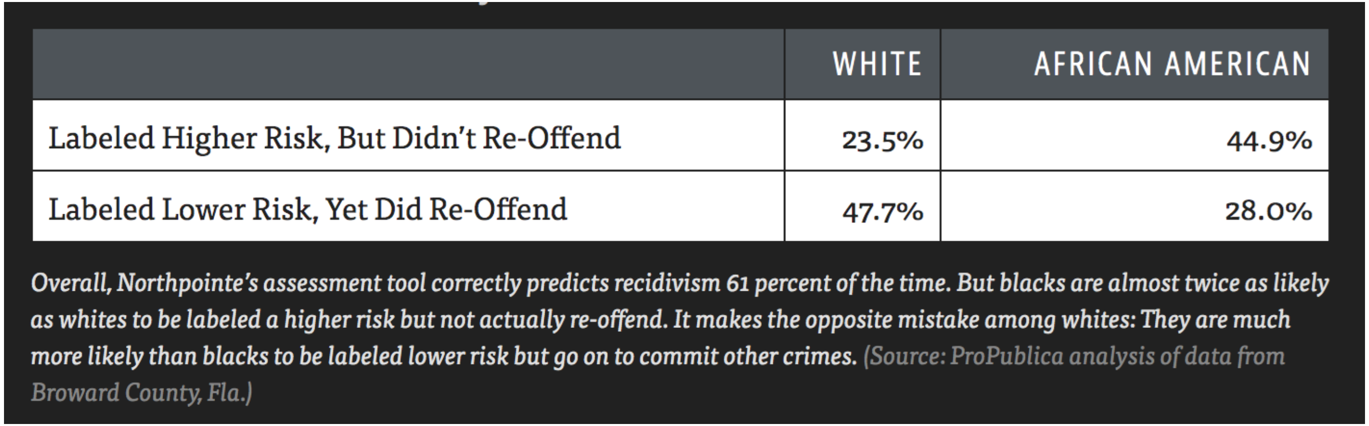


Machine Bias

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



	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%

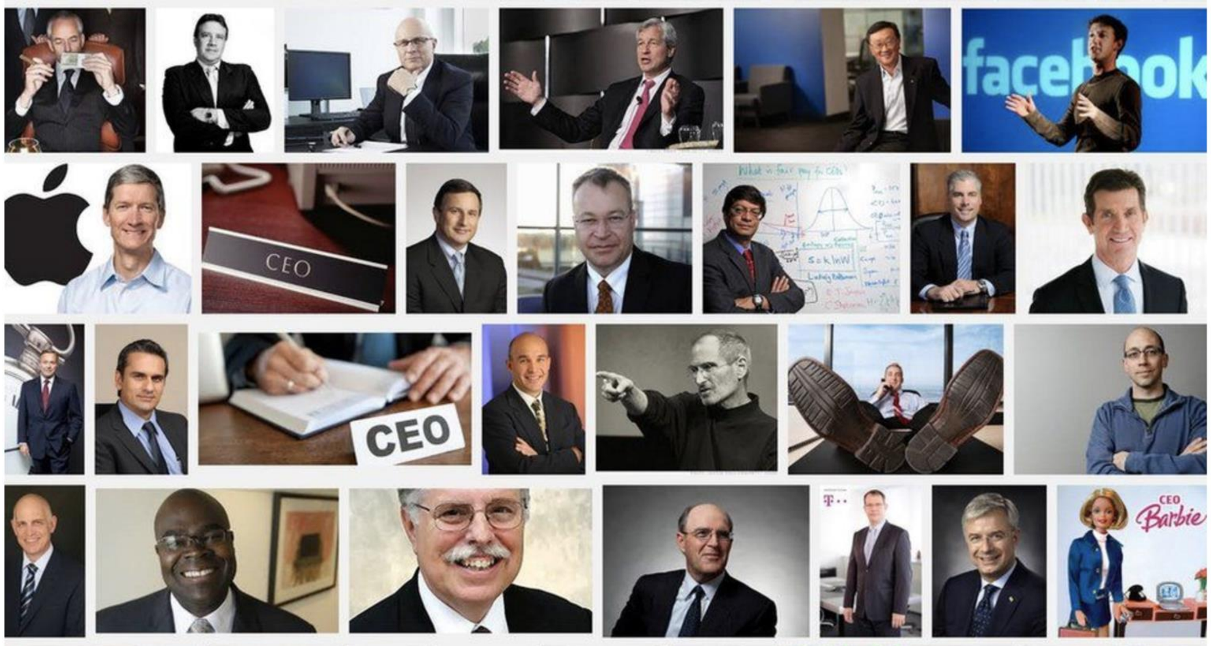
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Amazon Recruiting Algorithm

- Historically relatively few women got jobs at Amazon
- Penalized resumes with references to “women”
- Penalized graduates of women’s colleges

Representational harm: CEO Barbie?

Google search results for “CEO” used to show CEO Barbie as the first female image.



What is algorithmic bias?

- How the algorithmic decision-making process might systematically lead to worse outcomes for certain subpopulations
- Disparities that emerge due to demographic characteristics or other factors that are problematic from a societal perspective

Why is Legal Compatibility Important?

- To demonstrate evidence of algorithmic bias
  - Fairness metrics must conform to what judges, juries, or regulators would accept as evidence of discrimination or lack thereof
- To deploy bias mitigation methods
  - Methods themselves should not be discriminatory from a legal perspective

What are protected class variables?

- In U.S. law, protected classes are groups that are protected by anti-discrimination laws
- Examples of protected class variables include race, gender, age, disability, national origin, and religion
- Note that there are other variables of potential concern: e.g., socioeconomic class and geography
- Race and gender are treated symmetrically
  - E.g., white men can sue for discrimination on the basis of race or sex

Mitigating Algorithmic Bias

Why do we need to use protected class variables?

- “Fairness Through Unawareness:” Remove protected class variables and close proxies from training data
  - Widely considered “naïve” in the ML community
- Due to omitted variable bias, this doesn’t necessarily address bias
  - Combination of weak proxy variables can approximate protected class variables

Why do we need to use protected class variables?

- Most technical approaches to mitigating algorithmic bias require the use of protected class variables or proxies
    - Intuition: in order to mitigate disproportionate performance or outcomes across groups, need to take into account what groups people are in and actively reverse trends in data
- Including the protected class variable can sometimes reduce bias
    - Sometimes protected class variable can provide helpful context
        - E.g., if culture A steers best students to engineering and culture B steers best students to law, knowing which culture student is from is important if trying to predict scholastic ability

Yet "fairness through unawareness" is currently the most prevalent approach in industry.

Why is “fairness through unawareness” so common?

- Default if you don’t have access to protected class variables (often the case due to privacy concerns)
- On the surface, it is “blind” to protected classes
    - Often analogized to blindness in anti-discrimination law
- U.S. Department of Housing and Urban Development proposed rule for disparate impact would create a safe harbor for algorithms that do not use protected class variables or close proxies

Anti-Classification vs. Anti-Subordination

- Anti-Classification: classification or treatment that differs based on protected class attributes is discriminatory
- Anti-Subordination: law should seek to dismantle hierarchies between protected class groups even if doing so involves consciousness of these group classifications

The Court has increasingly adopted an anti-classification stance in recent years .

Anti-Discrimination Law

Relevant Anti-Discrimination Law

- Public Sector
    - Constitution, equal protection doctrine
    - Affirmative action doctrine (one of the few areas where the Court has permitted race-conscious decision-making)
- Private sector
    - Anti-discrimination statutes (e.g., Title VII of Civil Rights Act of 1964, Age Discrimination in Employment Act, Fair Housing Act)
    - Disparate impact vs. disparate treatment doctrines

What is Affirmative Action?

- Definition: preferential treatment for groups that have historically faced discrimination, usually in the context of school admissions, employment, or government contracting
- Also known as “positive discrimination” in the UK
- Historically applied to racial minorities and women in the United States

What are the connections between affirmative action and algorithmic fairness?

- Many of the proposals to rectify historical biases in algorithmic decision-making are forms of affirmative action.
- Affirmative action jurisprudence gives us a starting point from a legal perspective to evaluate potential measures to correct for biases.
- The controversies around affirmative action can give us an understanding of some of the consequences of implementing algorithmic affirmative action.
    - What are the issues we might be concerned about?
    - What are the controversies that might result?
    - What are the legal concerns?

Key Affirmative Action Cases

Regents of University of California v. Bakke (1978)

- Context
    - When University of California, Davis School of Medicine was founded in 1968, all-white class
    - School de-segregation was still a major issue across the U.S.
    - Reserved 16 out of 100 seats for admissions through a special committee
    - Goal of program was "to compensate victims of unjust societal discrimination”

Regents of University of California v. Bakke (1978)

- Decision
    - 9 Justices issued 6 opinions, plurality opinion written by Justice Powell
    - Racial quotas impermissible
    - Race can be one of many factors
    - KEY: Diversity is compelling state interest
        - Not redressing historical discrimination
            - Historical discrimination can only be a factor if there is evidence of the particular school discriminating (not just general societal discrimination)
- Connection to algorithmic fairness:
    - Usually trying to address disparities due to historical discrimination
    - Not focused on diversity
    - Often rebalancing

Grutter and Gratz v. Bollinger (2003)

- Two cases decided same day
- Context
  - Grutter:
    - University of Michigan law school argued that it had a compelling state interest to ensure a critical mass of minorities to obtain educational benefits of diversity
  - Gratz:
    - University of Michigan had 150-point scale for applicants: 100 points for admission, minorities gained additional 20 points

Grutter and Gratz v. Bollinger (2003)

- Decision
    - Law school’s system is permissible: Race can be a factor among many other factors evaluated on an individual basis (Grutter)
    - Undergraduate admissions system is impermissible: Point system not acceptable (Gratz) – too similar to quota
    - Students need to be individually assessed

How do we reconcile these rulings with algorithms?

- How can we have an algorithm that takes into account race but does not use quotas, different thresholds, or a point system?
- Justice Souter's dissent in Gratz:
    - "The very nature of a college's permissible practice of awarding value to racial diversity means that race must be considered in a way that increases some applicants' chances for admission. Since college admissions is not left entirely to inarticulate intuition, it is hard to see what is inappropriate in assigning some stated value to a relevant characteristic, whether it be reasoning ability, writing style, running speed, or minority race. Justice Powell's plus factors necessarily are assigned some values. The college simply does by a numbered scale what the law school accomplishes in its "holistic review," a distinction that does not imply that applicants to the undergraduate college are denied individualized consideration or a fair chance to compete on the basis of all of the various merits their applications may disclose."

Outside of Affirmative Action

Disparate Treatment

- Intentional discrimination
- Usually need proof of racism, sexism, etc.
  - E.g., boss makes derogatory remarks about minorities

Disparate Impact

- Does not require proof of intentionality
  - Worried about facially neutral policies that have unjustifiably disproportionate effects
- Burden-shifting framework:
  - Plaintiff shows disproportionate outcomes by group
  - Defendant shows business necessity
    - E.g., need firemen to be able to lift heavy weights, which might disproportionately make it harder for women to get job
  - Plaintiff shows there is less discriminatory option that achieves same business objectives

Texas v. Inclusive Communities (2015)

- Established disparate impact doctrine in housing context
- Plaintiff’s prima facie case must draw a **causal connection** between policy or practice and the statistical disparity
- HUD proposed rule based on this case
  - Safe harbor from disparate impact liability for algorithms that do not use protected class variables or close proxies
  - Has nothing to do with causality

Why does causality matter?

- Causality is important from a legal perspective: discrimination is defined as making a decision “because of X,” where “X” is a protected class variable
- Legal analysis around anti-discrimination liability involves questions of causality
- Judges will ultimately be evaluating causal relationships
    - E.g., would you or your model have made this same decision but for the individual’s race/sex/etc.?

Benefits of causal approaches

- Allows for distinctions to be made between different uses of protected class variables
    - Are you trying to be biased or mitigate bias?
    - Are you increasing or decreasing the causal relationship between the protected class attribute and the model predictions/decisions?
- Causation vs. correlation: not all correlations with protected class attributes are harmful (context is important)

Takeaways

- Legal compatibility is key to employing algorithmic bias mitigation techniques in practice
- There is a tension between the technical need to consider protected class attributes in order to mitigate bias and the law's preference for decision-making that is blind or neutral to these attributes
- Causality is a key concept in both ML and law that can help distinguish between different uses of protected class variables
  - Hopefully make it possible to prevent proliferation of biased algorithms and permit use of bias mitigation techniques

Thanks for listening!

Questions?

These topics are discussed further in my forthcoming law review article, “Reconciling Legal and Technical Approaches to Algorithmic Bias,” to be published in the *Tennessee Law Review*.