Modern software influences critical decisions

Software Fairness
April 11, 2023


The Government Is Blacklisting People Based on Predictions of Future Crimes

On Orbitz, Mac Users Steered To Pricier Hotels

Minority homebuyers face widespread statistical lending discrimination, study finds

The Algorithm That Beats Your Bank Manager
Software can make bad decisions.
Software can discriminate!
YouTube automatic captions

Rachael Tatman, “Gender and Dialect Bias in YouTube’s Automatic Captions” in 2017 Workshop on Ethics in Natural Language Processing

Joy Buolamwini
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

how people want to use vision software

today’s goals

Define software discrimination.
Operationalize measuring discrimination through causal software testing.
Provide provable fairness guarantees.
Design software to be fair

Typically machine learning systems:
- Balance training sets
- Introduce training noise
- Constrain regression’s loss function
- Split criteria on sensitive inputs

Design alone is not enough

Possible causes:
- Biased data
- Implementation bugs
- Unintended interactions and mismatched components

Fairness must be part of the software engineering lifecycle

Fairness is just like quality and security

Call to Action! 
Fairness must be part of the software engineering lifecycle

Requirements

Call to Action! 
We need methods for specifying fairness requirements

Call to Action! 
We need fairness design principles
Let’s talk about what it means for systems to discriminate.

**Requirements**

**Design**

**Testing**

**Verification**

We need automated fairness testing.

We need fairness property verification.

Fairness must be part of the software engineering lifecycle.

**LOAN program**

This talk is not about policy.

**Fairness: Disparate Treatment**

**Fairness: Disparate Treatment**

- **Hide the data**

  - Ads by Google

  - Latanya Sweeney, Arrested?
    - 1) Enter Name and State. 2) Access Full Background Checks Instantly.
    - www.instantcheckmate.com/

  - Ineffective because of data correlation.
    - [Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

**Fairness: Demographic Parity**

- **Compare subpopulation proportions**

  - Often called group discrimination

  - Fails to identify discrimination against individuals.

**How group discrimination can fail**

- **Europe**
  - Approve loans to all *green* deny loans to all *purple* applicants

- **Asia**
  - Approve loans to all *purple* deny loans to all *green* applicants

- European and Asian discriminations cancel each other out, and the group discrimination measure can be 0.

**Fairness: Disparate Impact**

- **Prohibits using a facially neutral practice that has an unjustified adverse impact on members of a protected class.**

- 80% rule: Employer's hiring rates for protected groups may not differ by more than 80%.

- Zafar et al. Fairness constraints: Mechanisms for fair classification. AISTATS 2017

---

- **Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women**


- Disparate treatment: still not fair

- Compare subpopulation proportions

- Fails to identify discrimination against individuals.

- 35% 65%

- 80% 20%

- European and Asian discriminations cancel each other out, and the group discrimination measure can be 0.
Fairness: Delayed Impact

Making seemingly fair decisions can (but shouldn’t), in the long term, produce unfair consequences.

Liu et al., Delayed impact of fair machine learning. ICML 2018

Fairness: Predictive Equality

False positive rates should not differ

Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. FATML 2016

Corbett-Davies. Algorithmic decision making and the cost of fairness. ICML 2017

Fairness: Equal Opportunity

False negative rates should not differ


Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments FATML 2016

Fairness: Equalized Odds

Predictive equality and equal


Fairness: Treatment Equality

Consistent ratios of false positives to false negatives for each group

Berk et al. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research 2018

Fairness: Correlation

correlation(race, score) = 0.8

mutual information(race, score) = 0.6

Correlation does not measure causation


What is fairness?

Sensitive inputs should not affect software behavior.

We want to measure causality!

Sensitive inputs should not affect software behavior.

No need for an oracle!

Themis generates a test suite or can use a manually written one.

http://fairness.cs.umass.edu

Apparent discrimination can be group or causal, measured on a given test suite or operational profile.
How does Themis work? adaptive, confidence-driven sampling

input schema
certainty
error bound

sound pruning

race race race age age gender gender location

findings

Trying to avoid group discrimination

Training a decision tree not to discriminate against gender made it discriminate against race 38.4% of the time.

evaluation

Eight open-source decision systems trained on two public data sets

| Discrimination-aware logistic regression | Pruning
| Discrimination-aware decision tree | 148×
| Discrimination-aware naive Bayes | 2,849×
| Discrimination-aware decision tree | 13,000×
| Naive Bayes | scikit-learn
| Decision tree |
| Logistic regression |

findings

Group discrimination is not enough.

More than 11% of the individuals had their output flipped just by altering the individual’s gender.

Decisio tree trained not to group discriminate against gender causal discriminated against gender: 0.11.

findings

Pruning is highly effective.

- The more a system discriminates, the more efficient Themis is.
- On average, pruning reduced test suites by 148× for causal and 2,849× for group discrimination. Best improvement was 13,000×.