Bias in Software Systems

Yuriy Brun

https://laser.cs.umass.edu/

Modern software influences critical decisions
Wisconsin Supreme Court allows state to continue using computer program to assist in sentencing

Software can make bad decisions. Software can discriminate!
YouTube automatic captions

Oh Jessica I am this stove I play the heroine me I am
today's goals

Define software discrimination.

Operationalize measuring discrimination through causal software testing.

Provide provable fairness guarantees.

possible causes

- biased data
- implementation bugs
- unintended interactions and mismatched components

Design alone is not enough

Let's talk about what it means for systems to discriminate
This talk is not about policy.

Ineffective because of data correlation.

[Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

[Latanya Sweeney. Arrested?] 1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

[Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]


Disparate treatment: still not fair
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

Asia
- Approve loans to all purple
- Deny loans to all green

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%.

Fairness: Delayed Impact

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

Fairness: Predictive Equality

- False positive rates should not differ

Fairness: Equal Opportunity

- False negative rates should not differ

Fairness: Equalized Odds

- Predictive equality and equal

Fairness: Treatment Equality

- Consistent ratios of false positives to false negatives for each group

Metric Fairness

- Representation Disparity
- Conditional Use Accuracy
- Accuracy Equality
Correlation does not measure causation

**Fairness: Correlation**

\[
\text{correlation}(\text{race}, \text{approved}) = 0.8
\]

\[
\text{mutual information}(\text{race}, \text{approved}) = 0.6
\]

Correlation does not measure causation

**What is fairness?**

Sensitive inputs should not affect software behavior.

We want to measure causality!

---

**causal testing**

Sensitive inputs should not affect software behavior.

No need for an oracle!
Themis generates a test suite or can use a manually written one

How much does my software discriminate with respect to …?

Does my software discriminate more than 10% of the time, and against

http://fairness.cs.umass.edu

How much does my software discriminate with respect to …?

Does my software discriminate more than 10% of the time, and against

Themis


Causal discrimination

LOAN( ) \neq LOAN ( )

Group discrimination

\begin{array}{l}
\text{Census income dataset: financial data} \\
\text{45K people income > $50K?} \\
\text{Statlog German credit dataset: credit data} \\
\text{1K people "good" or "bad" credit?}
\end{array}

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

Group discrimination is not enough.

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

Trying to avoid group discrimination

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

Apparent discrimination can be group or causal, measured on a given test suite or operational profile.
Debugging

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

*The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among them civil rights legend Rep. John Lewis (D-Ga.).*
What are we doing now?

- Fair computer vision
- Fair natural language processing

But what’s the holy grail?

- Provably fair machine learning: Provide (high-probability) guarantees that the classifier is fair on unseen data.

Disparate Impact


Demographic Parity

Equal Opportunity

Equalized Odds

Predictive Equality
Contributions

- **Causality-based** definition and method for measuring software fairness
- Themis, an automated test-suite generator for fairness testing
- Evaluation on real-world software, demonstrating software is biased and our methods can catch it
- Provable guarantees on fairness in machine learning

http://fairness.cs.umass.edu

http://fairness.cs.umass.edu

https://tinyurl.com/FairnessPaper

Yuriy Brun
Alexandra Meliou
Sainyam Galhotra
Rico Angel
Brittany Johnson
Stephen Giguer
Saket Brockens
Maorin Michiner
Sahyoun Ghassemi

Contributions

- **Causality-based** definition and method for measuring software fairness
- Themis, an automated test-suite generator for fairness testing
- Evaluation on real-world software, demonstrating software is biased and our methods can catch it
- Provable guarantees on fairness in machine learning

http://fairness.cs.umass.edu

http://fairness.cs.umass.edu

https://tinyurl.com/FairnessPaper

UMass Amherst

ORACLE