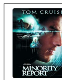
Bias in Software Systems

Yuriy Brun

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



Resilient cities Cities

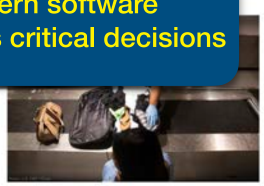
Predicting crime, LAPD-style

Cutting edge data-driven analysis directs Los Angeles patrol officers to likely future crime scenes - but critics worry that decision-making by machine will bring 'tyranny of the algorithm'

● Join our live Q&A with Homicide Watch this Friday

▲ PredPol co-developer P Jeffrey Brantingham at the Unified Command Post in Los Angeles. 'This is not Minority Report,' he said. Photograph: Damian Dovarganes/AP

https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report



ACLU GET UPDATES / DONATE

The Government Is Blacklisting People Based on Predictions of Future Crimes

By Hina Shamsi, Director, ACLU National Security Project

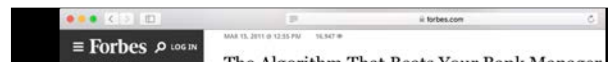
Modern software influences critical decisions

https://www.aclu.org/blog/national-security/discriminatory-profiling/government-blacklisting-people-based-predictions

THE WALL STREET JOURNAL.

On Orbitz, Mac Users Steered to Pricier Hotels

By Dana Mattioli



Forbes

The Algorithm That Beats Your Bank Manager

HAAS NEWS > NEWS CATEGORIES > RESEARCH NEWS

Minority homebuyers face widespread statistical lending discrimination, study finds

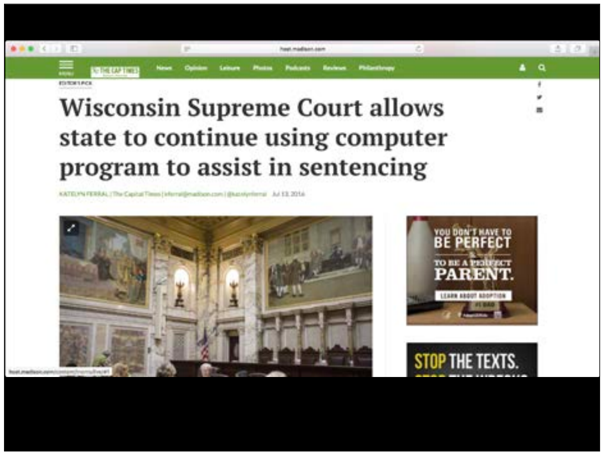
By Laura Counts | NOVEMBER 13, 2018

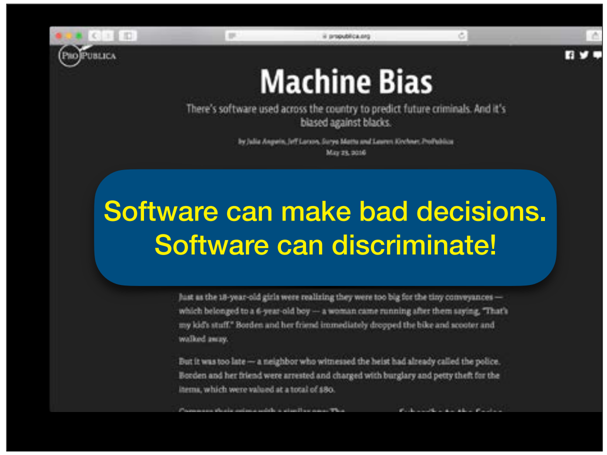
Face-to-face meetings between mortgage officers and homebuyers have been rapidly replaced by online applications and algorithms, but lending discrimination hasn't gone away.

A new University of California, Berkeley study has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers, earning 11 to 17 percent higher profits on such loans. All told, those homebuyers pay up to half a billion dollars more in interest every year than white borrowers with comparable credit scores do, researchers found.

The findings raise legal questions about the rise of statistical discrimination in the fintech era, and point to potentially widespread violations of U.S. fair lending laws, the researchers say. While lending discrimination has historically been caused by human prejudice, pricing disparities are increasingly the result of algorithms that use machine learning to target applicants who might shop around less for higher-priced loans.

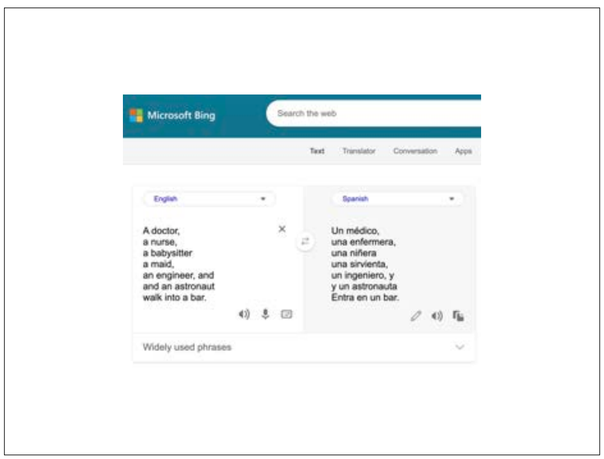
"The mode of lending discrimination has shifted from human bias to algorithmic bias," said study co-author Adair Morse, a finance professor at UC Berkeley's Haas School of Business. "Even if the people writing the

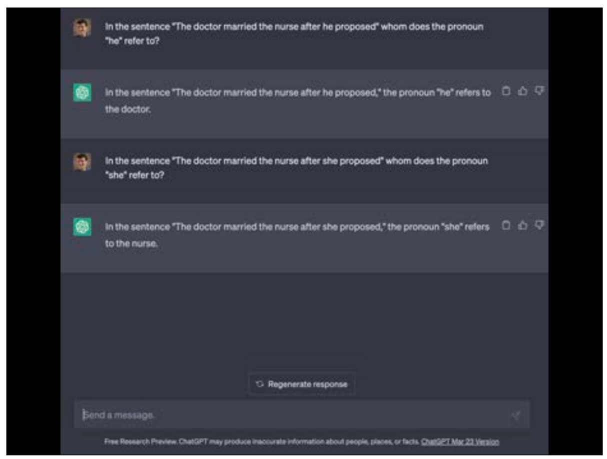




Google Translate

BING TRANSLATOR

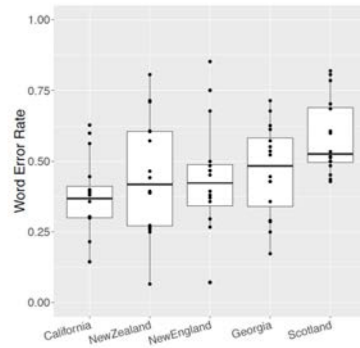




YouTube

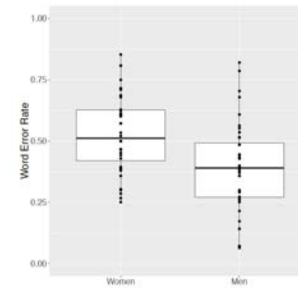
YouTube automatic captions

YouTube automatic captions

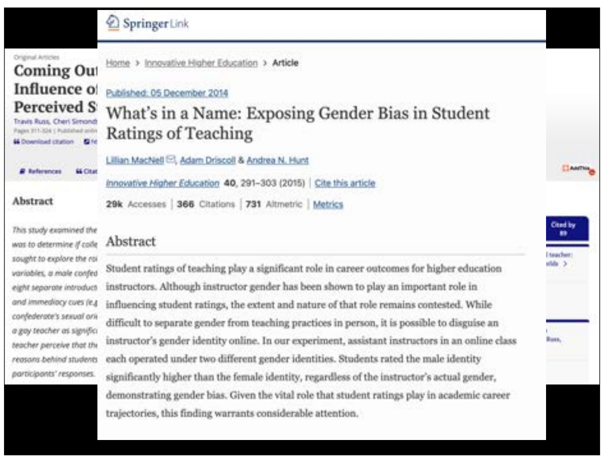


Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions" in 2017 Workshop on Ethics in Natural Language Processing

YouTube automatic captions



Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions" in 2017 Workshop on Ethics in Natural Language Processing



SpringerLink

Home > Innovative Higher Education > Article

Published: 05 December 2014

What's in a Name: Exposing Gender Bias in Student Ratings of Teaching

Lillian MacNell, Adam Driscoll & Andrea N. Hunt

Innovative Higher Education 40, 291–303 (2015) | Cite this article

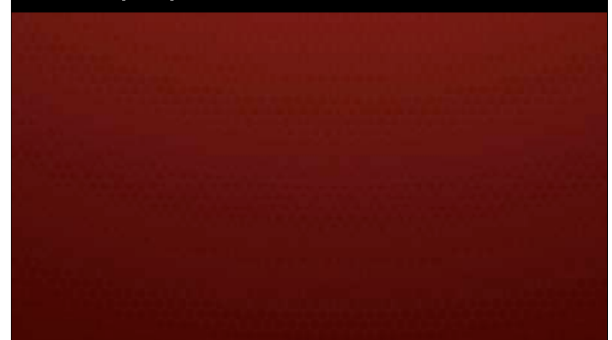
29k Accesses | 366 Citations | 731 Altmetric | Metrics

Abstract

Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention.

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

how people want to use vision software



how people want to use vision software

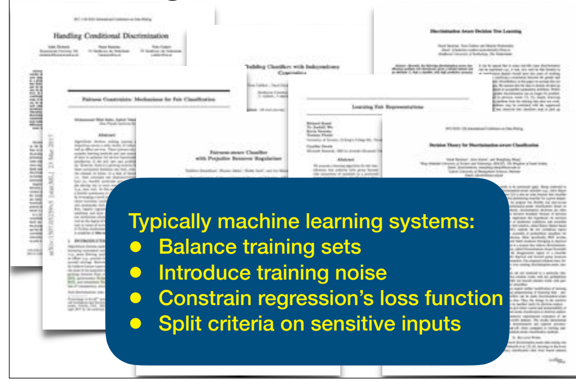
today's goals

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

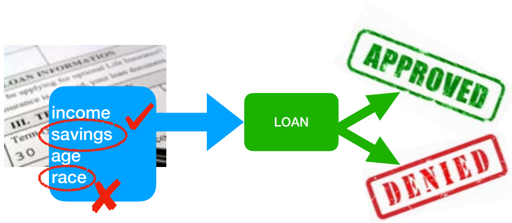
poor design

implementation bugs

unintended interactions and mismatched components

Let's talk about what it means for systems to discriminate

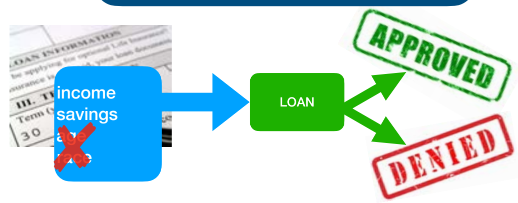
LOAN program



This talk is not about policy.

Fairness: Disparate Treatment

Hide the data

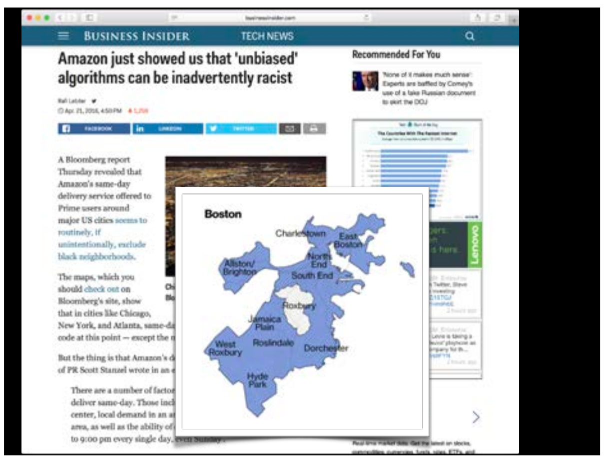


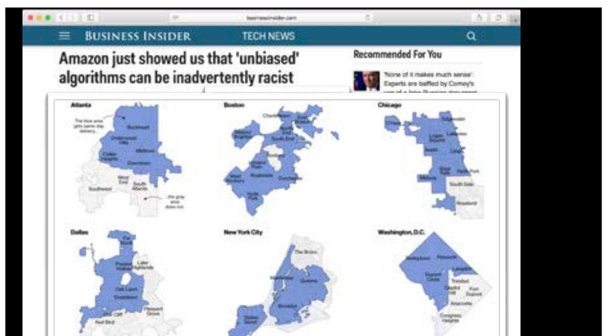
Zafar et al. Fairness constraints: Mechanisms for fair classification. AISTATS 2017.

Fairness: Disparate Treatment

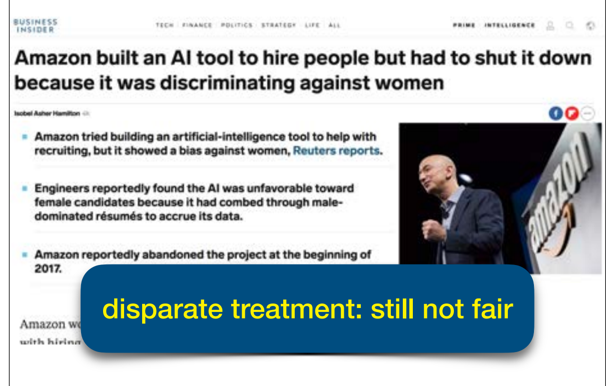
Hide the data

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





https://www.bloomberg.com/graphics/2016-amazon-same-day/



disparate treatment: still not fair

https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

Fairness: Demographic Parity

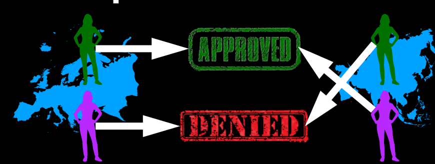
Compare subpopulation proportions

often called group discrimination

Fails to identify discrimination against individuals.

Dwork et al. Fairness through awareness. ITCS 2012.
Calders and Verwer. Three naive Bayes approaches for discrimination-free classification. DMKD 2010.

How group discrimination can fail



approve loans to all green deny loans to all purple applicants

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.

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. KDD 2017

Fairness: Equal Opportunity

False negative rates should not differ

Hardt et al. Equality of Opportunity in Supervised Learning. NIPS 2016
Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments FATML 2016

Fairness: Equal… Odds

…lictive eq…

…earning. NIPS 2016

Use Accuracy

…rness

Accuracy Equality

…ation Disparity

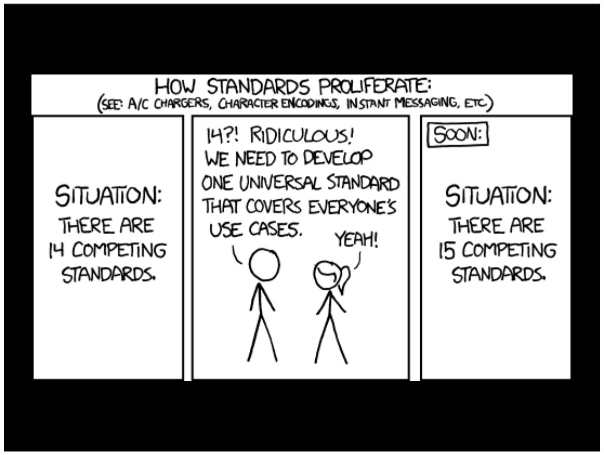
Conditi…

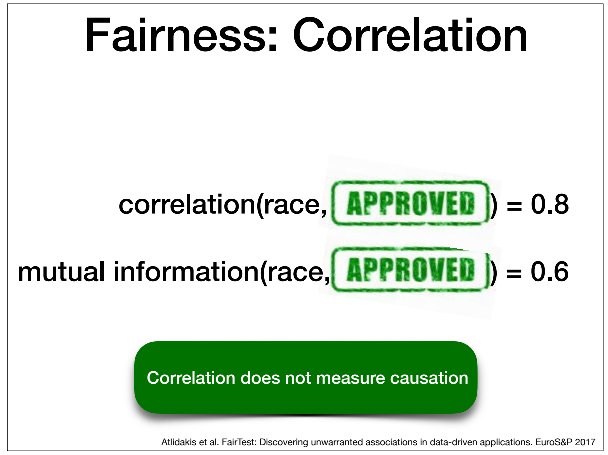
Fairn… Equality

…to

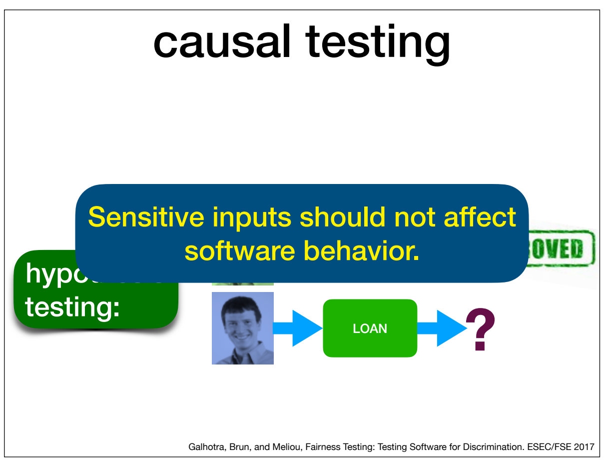
…eg…

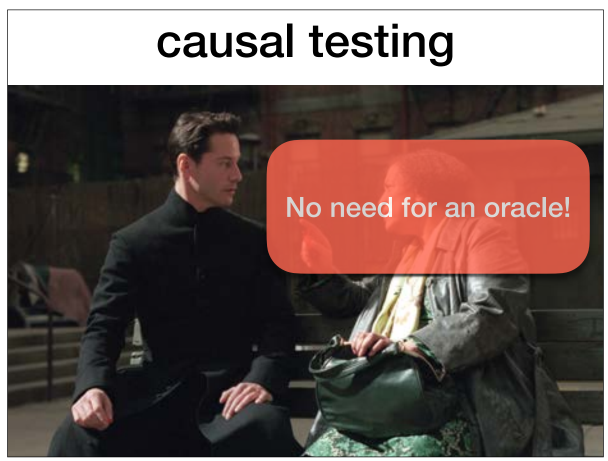
Berk et al. Fairness in criminal justice … rt. Sociol… 18

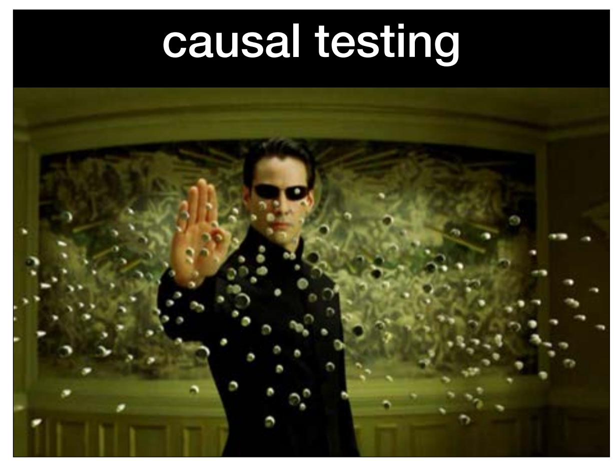

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.


Fairness: Correlation
correlation(race, APPROVED) = 0.8
mutual information(race, APPROVED) = 0.6
Correlation does not measure causation
Atlidakis et al. FairTest: Discovering unwarranted associations in data-driven applications. EuroS&P 2017


What is fairness?
Sensitive inputs should not affect software behavior.
We want to measure causality!
Judea Pearl. Causal inference in statistics: An overview. Statistics Surveys 2009


causal testing
Sensitive inputs should not affect software behavior.
testing:
LOAN
?
Galhotra, Brun, and Meliou, Fairness Testing: Testing Software for Discrimination. ESEC/FSE 2017


causal testing
No need for an oracle!


causal testing

Themis

automated test-suite generator

How much does my software discriminate with respect to …?

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

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

http://fairness.cs.umass.edu

Angell, Johnson, Brun, and Meliou, Themis: Automatically Testing Software for Discrimination. ESEC/FSE 2018 Demo

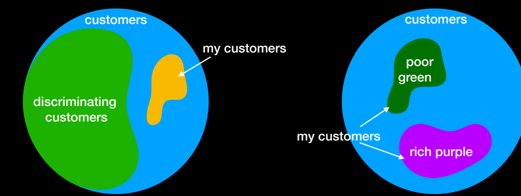
discrimination measures

causal discrimination

LOAN( ) $\stackrel{?}{=}$ LOAN ( )

group discrimination

apparent discrimination



Software may discriminate, but not for a given set of customers

Fair software may appear to discriminate
(e.g., Amazon same-day delivery)

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

Evaluation

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

discrimination-aware logistic regression	[88]
discrimination-aware decision tree	[40]
discrimination-aware naive Bayes	[18]
discrimination-aware decision tree	[91]
naive Bayes	scikit-learn
decision tree	
logistic regression	
SVM	

- Census income dataset:
  financial data
  45K people
  income > $50K?
- Statlog German credit dataset:
  credit data
  1K people
  “good” or “bad” credit?

findings

Group discrimination is not enough.

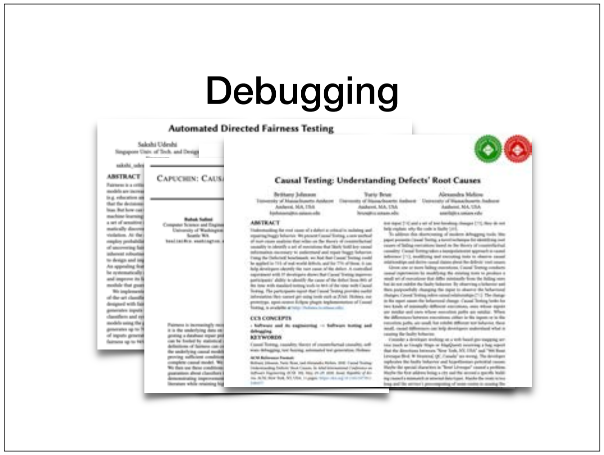
More than 11% of the individuals had the output flipped just by altering the individual’s gender.

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

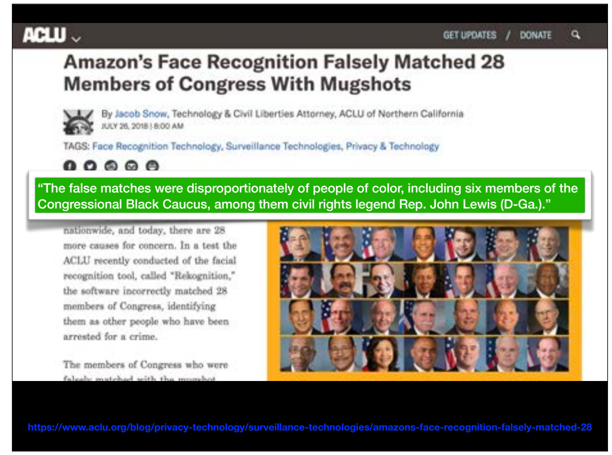
findings

Trying to avoid group discrimination

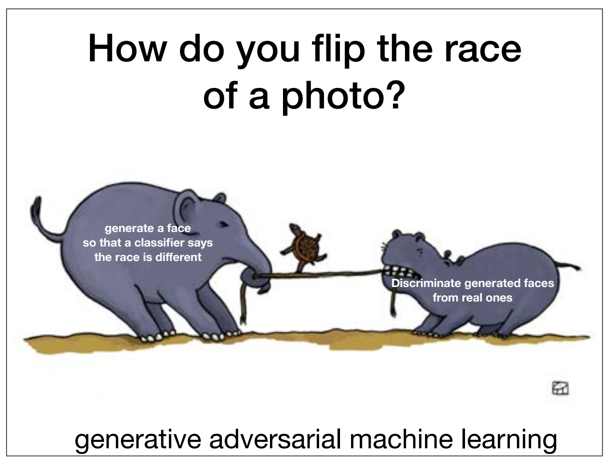
Training a decision tree not to discriminate against gender made it discriminate against race 38.4% of the time.


Debugging
Automated Directed Fairness Testing
Causal Testing: Understanding Defects' Root Causes


fairkit-learn
Fairkit, Fairkit, on the Wall, Who's the Fairest of Them All?
Supporting Data Scientists in Training Fair Models
Brittany Johnson, Jesse Bartola, Rico Angell, Katherine Keith,
Sam Witty, Stephen J. Giguere, Yuriy Brun


ACLU
GET UPDATES / DONATE
Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots
By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM
TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology
"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.)."
nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime.
The members of Congress who were
https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28


Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots
Fair computer vision
https://thispersondoesnotexist.com/

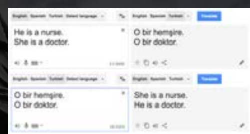

How do you flip the race of a photo?
generate a face so that a classifier says the race is different
Discriminate generated faces from real ones
generative adversarial machine learning


this-person-does-not-exist.com

What are we doing now?

Fair computer vision

Fair natural language processing



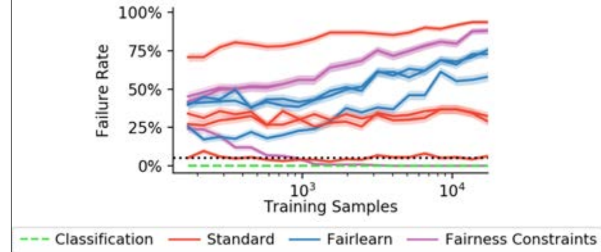
applied

But what's the holy grail?

Provably fair machine learning:

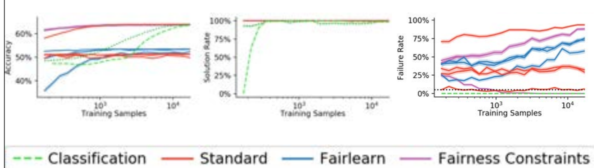
Provide (high-probability) guarantees that the classifier is fair on unseen data.

Disparate Impact



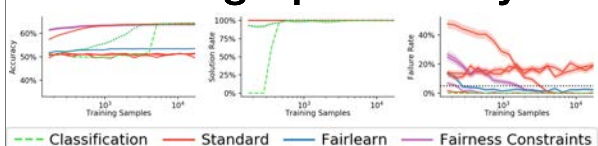
Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

Disparate Impact

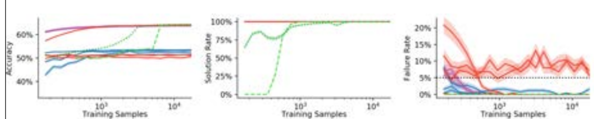


Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

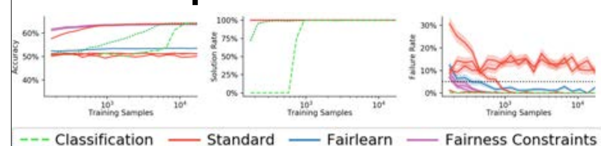
Demographic Parity



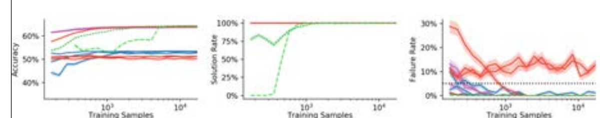
Equal Opportunity



Equalized Odds



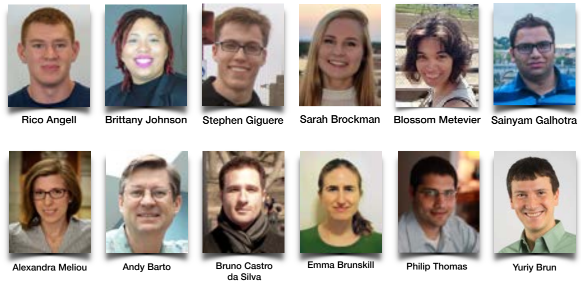
Predictive Equality



Contributions

http://fairness.cs.umass.edu

- 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

https://tinyurl.com/FairnessPaper

NSF UMassAmherst ORACLE

Contributions

http://fairness.cs.umass.edu

- 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