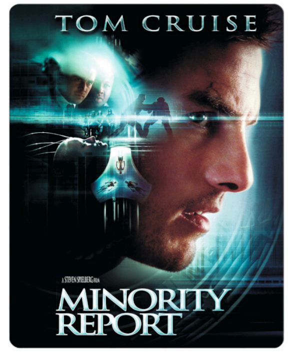
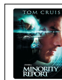
Bias in Software Systems

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

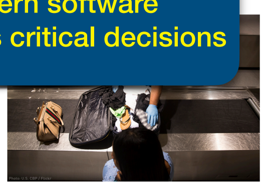
The Government Is Blacklisting People Based on Predictions of Future Crimes

By Hina Shamsi, Director, ACLU National Security Project

Modern software influences critical decisions

funerals or religious obligations, and lose jobs because you can't travel or your employer finds out you're blacklisted.

You know what the government has done violates your constitutionally protected ability to travel and to be free from false stigma. You have



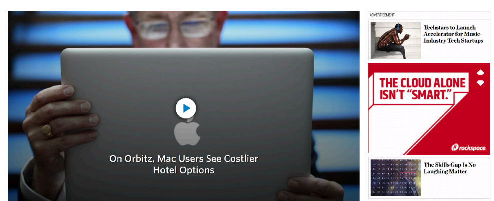
https://www.aclu.org/blog/national-security/discriminatory-profiling/government-blacklisting-people-based-predictions

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

TECH

On Orbitz, Mac Users Steered to Pricier Hotels



By Dana Mattioli

≡ Forbes LOG IN

The Algorithm That Beats Your Bank Manager

HAAS NEWS > NEWS CATEGORIES > RESEARCH NEWS

Minority homebuyers face widespread statistical lending discrimination, study finds

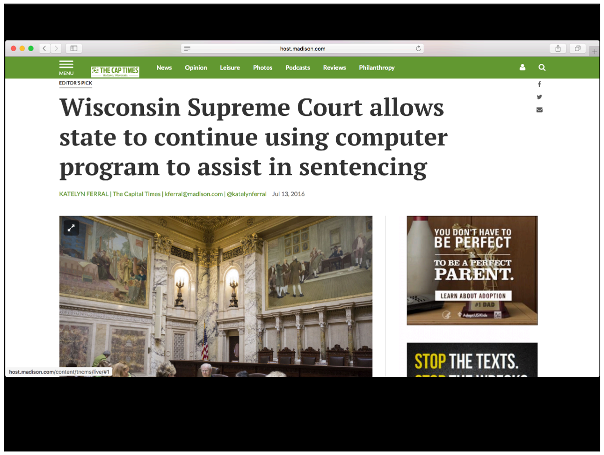
By Laura Counts | NOVEMBER 13, 2018

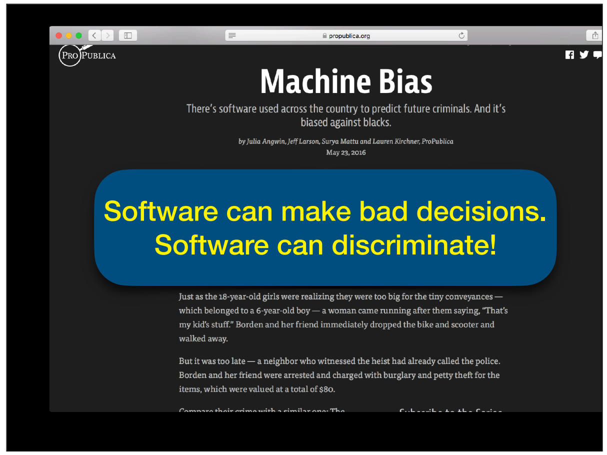
Face-to-face meetings between mortgage officers and homebuyers have been rapidly replaced by online applications and algorithms, but lending discrimination hasn't gone away.

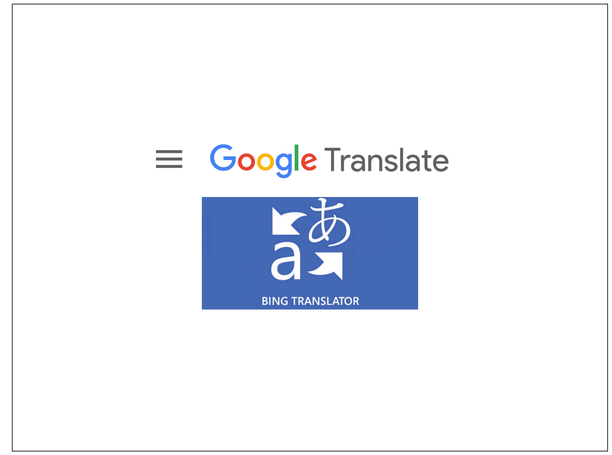
A new University of California, Berkeley study has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers, earning 11 to 17 percent higher profits on such loans. All told, those homebuyers pay up to half a billion dollars more in interest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

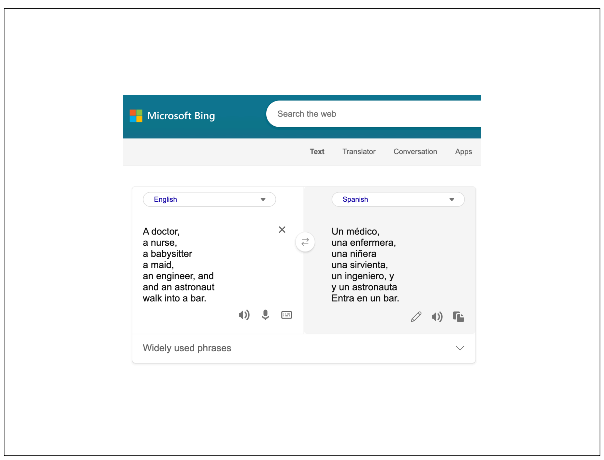
The findings raise legal questions about the rise of statistical discrimination in the fintech era, and point to potentially widespread violations of U.S. fair lending laws, the researchers say. While lending discrimination has historically been caused by human prejudice, pricing disparities are increasingly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

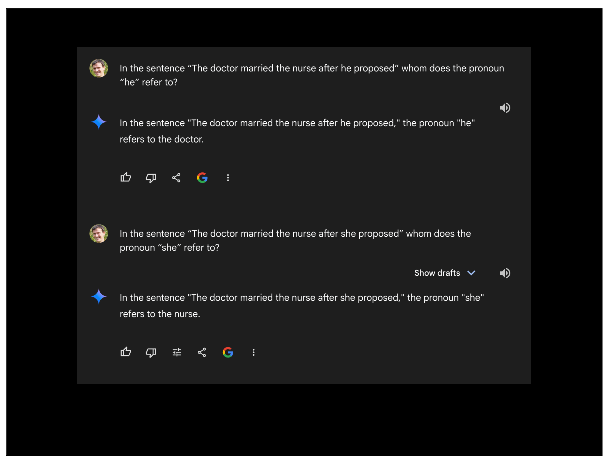
"The mode of lending discrimination has shifted from human bias to algorithmic bias," said study co-author Adair Morse, a finance professor at UC Berkeley's Haas School of Business. "Even if the people writing the

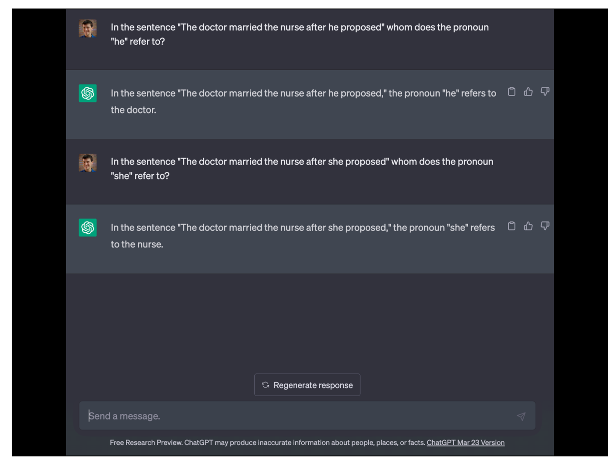

News
Opinion
Leisure
Photos
Podcasts
Reviews
Philanthropy
Wisconsin Supreme Court allows state to continue using computer program to assist in sentencing
YOU DON'T HAVE TO BE PERFECT TO BE A PERFECT PARENT.
LEARN ABOUT ADOPTION
STOP THE TEXTS.


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
Software can make bad decisions.
Software can discriminate!
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.
But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

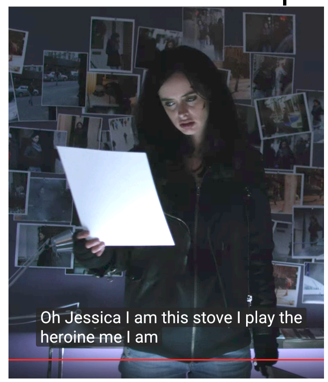

Google Translate
BING TRANSLATOR


Microsoft Bing
Search the web
Text
Translator
Conversation
Apps
English
Spanish
A doctor,
a nurse,
a babysitter
a maid,
an engineer, and
and an astronaut
walk into a bar.
Un médico,
una enfermera,
una niñera
una sirvienta,
un ingeniero, y
y un astronauta
Entra en un bar.
Widely used phrases

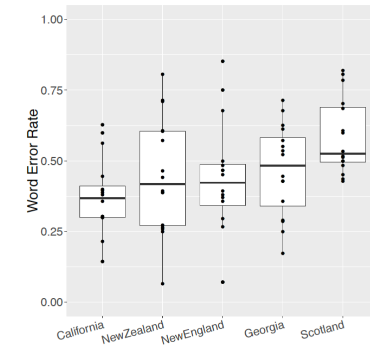

In the sentence "The doctor married the nurse after he proposed" whom does the pronoun "he" refer to?
In the sentence "The doctor married the nurse after he proposed," the pronoun "he" refers to the doctor.
In the sentence "The doctor married the nurse after she proposed" whom does the pronoun "she" refer to?
Show drafts
In the sentence "The doctor married the nurse after she proposed," the pronoun "she" refers to the nurse.


In the sentence "The doctor married the nurse after he proposed" whom does the pronoun "he" refer to?
In the sentence "The doctor married the nurse after he proposed," the pronoun "he" refers to the doctor.
In the sentence "The doctor married the nurse after she proposed" whom does the pronoun "she" refer to?
In the sentence "The doctor married the nurse after she proposed," the pronoun "she" refers to the nurse.
Regenerate response
Send a message.
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT Mar 23 Version

YouTube automatic captions

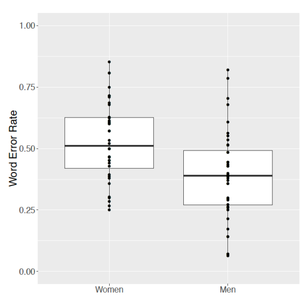


YouTube automatic captions



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

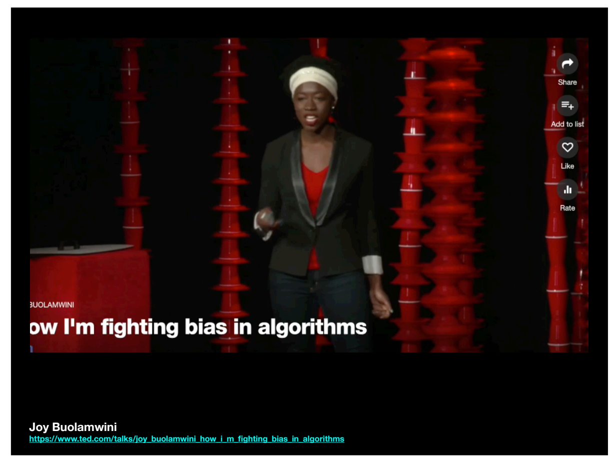
Lillian MacNell, Adam Driscoll & Andrea N. Hunt

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

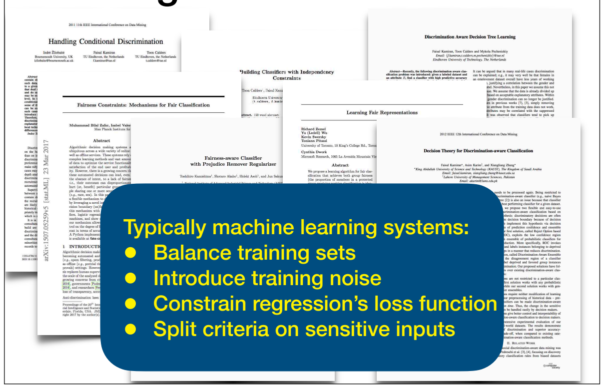
today's goals

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

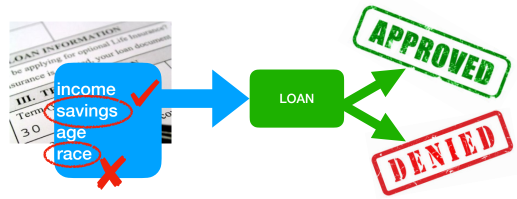
poor design

implementation bugs

unintended interactions and mismatched components

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

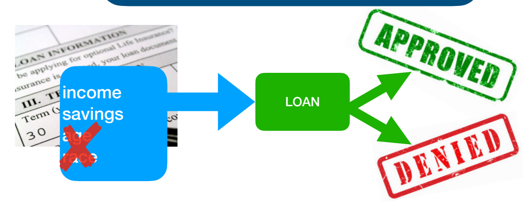
LOAN program



This talk is not about policy.

Fairness: Disparate Treatment

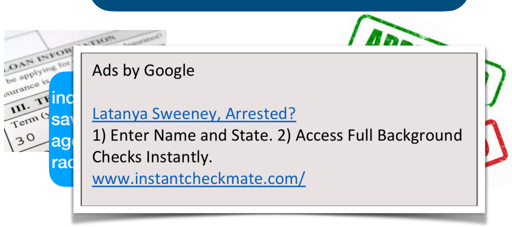
Hide the data



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

Fairness: Disparate Treatment

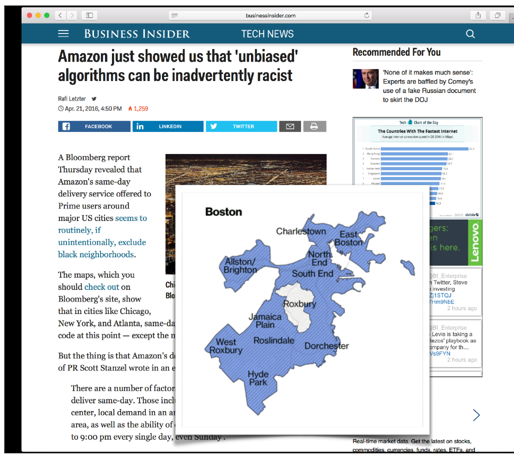
Hide the data

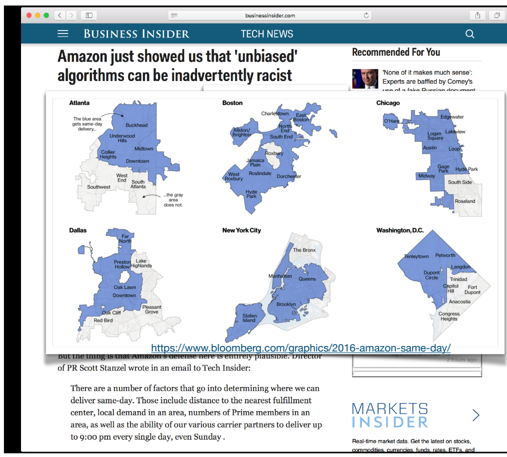


Ads by Google

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]





BUSINESS INSIDER

TECH · FINANCE · POLITICS · STRATEGY · LIFE · ALL   PRIME · INTELLIGENCE

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

Isobel Asher Hamilton 4h

- Amazon tried building an artificial-intelligence tool to help with recruiting, but it showed a bias against women, Reuters reports.
- Engineers reportedly found the AI was unfavorable toward female candidates because it had combed through male-dominated résumés to accrue its data.
- Amazon reportedly abandoned the project at the beginning of 2017.

Amazon w... with hiring...

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

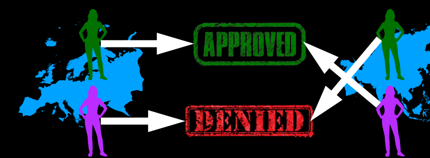
often called group discrimination

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

Europe   Asia



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

Fairn... Equality

...to

...ega...

Berk et al. Fairness in criminal justice ... rt. Sociolo... 18

Use Accuracy

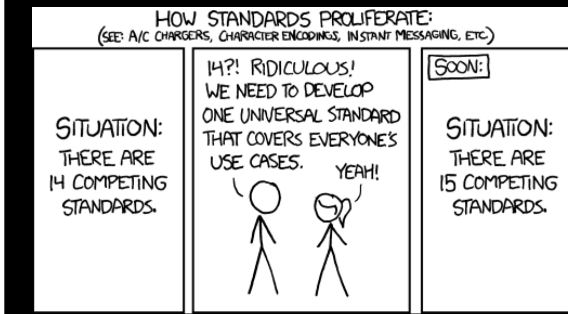
...rness

...orre...

Accuracy Equality

...ation Disparity

Conditi...



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!

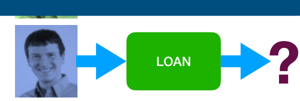
Judea Pearl. Causal inference in statistics: An overview. Statistics Surveys 2009

causal testing

Sensitive inputs should not affect software behavior.

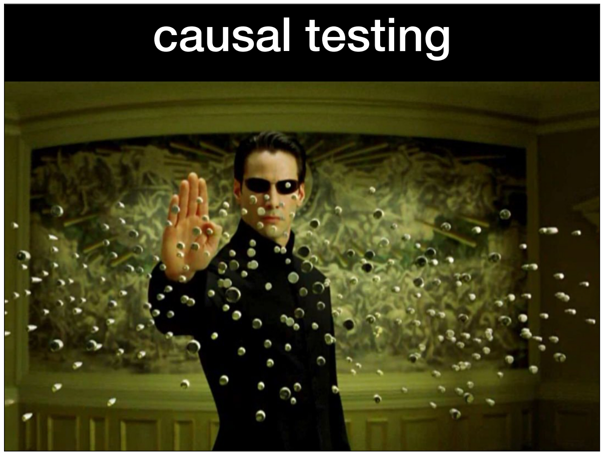
hypo... testing:

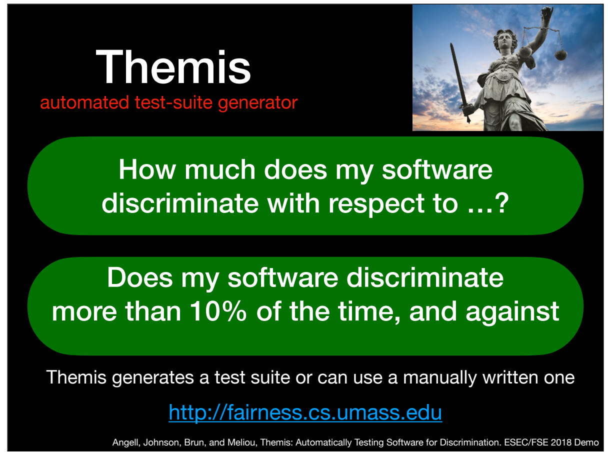
...OVED

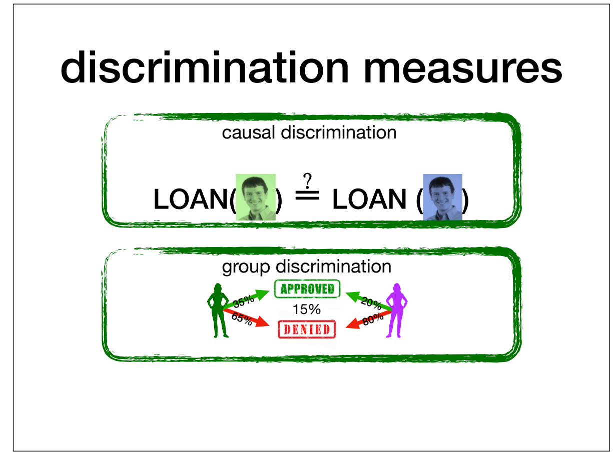


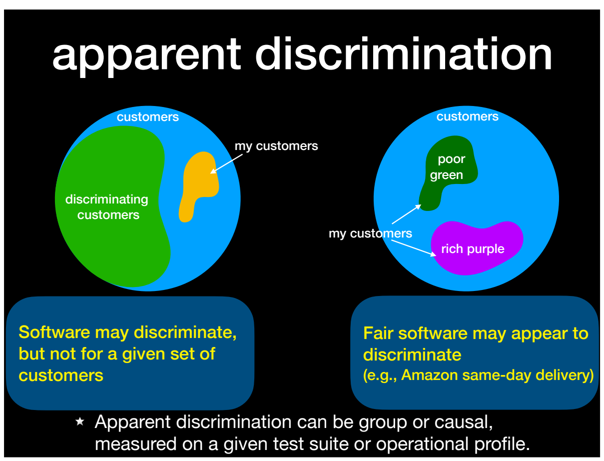
Galhotra, Brun, and Meliou, Fairness Testing: Testing Software for Discrimination. ESEC/FSE 2017

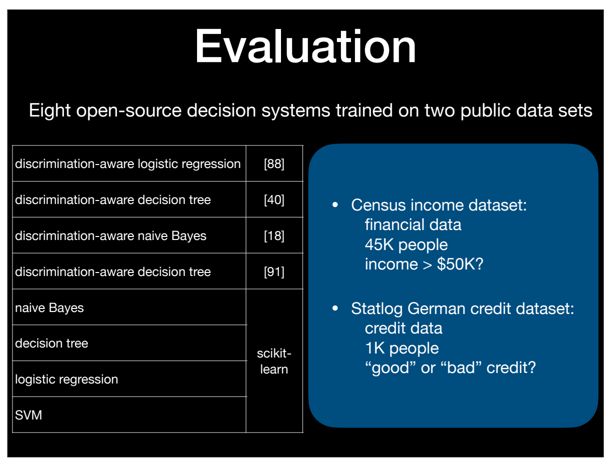
causal testing


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( ) ≟ LOAN ( )
group discrimination
APPROVED
15%
DENIED


apparent discrimination
customers
my customers
discriminating customers
customers
poor green
my customers
rich purple
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
decision tree
logistic regression
SVM
scikit-learn
• Census income dataset:
financial data
45K people
income > $50K?
• Statlog German credit dataset:
credit data
1K people
"good" or "bad" credit?

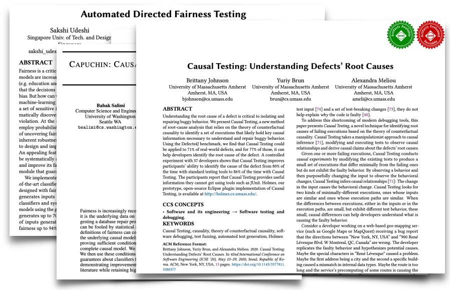

findings
Group discrimination is not enough.
More than 11% of the individuals had the output flipped just by altering the individual's gender.
Decision tree trained not to group discriminate against gender causal discriminated against gender: 0.11.

findings

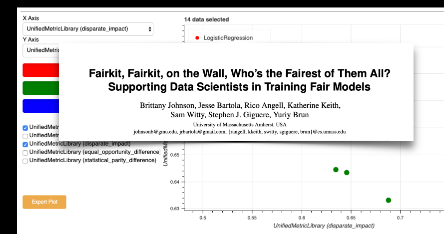
Trying to avoid group discrimination

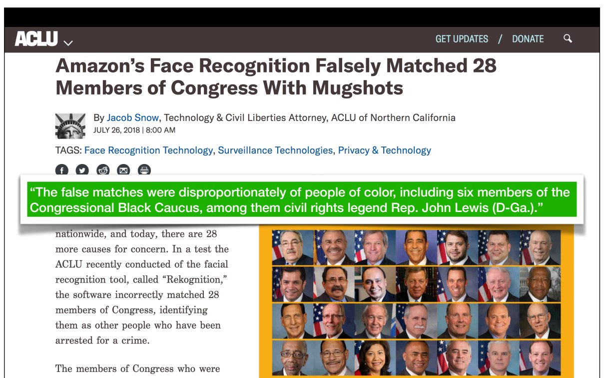
Training a decision tree not to discriminate against gender made it discriminate against race 38.4% of the time.

Debugging



fairkit-learn





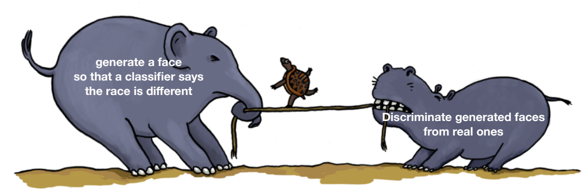
"The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among them civil rights legend Rep. John Lewis (D-Ga.)."

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Fair computer vision

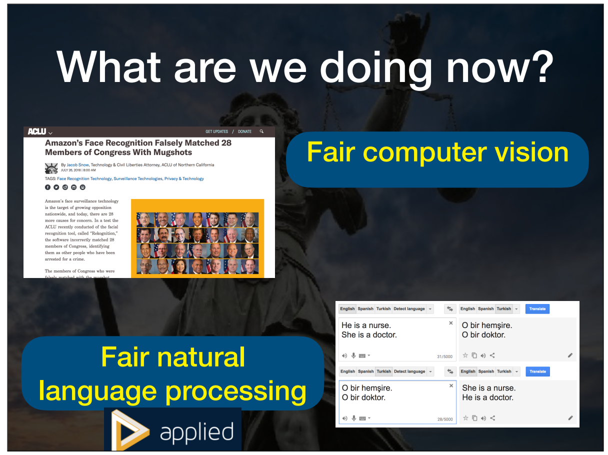
https://thispersondoesnotexist.com/

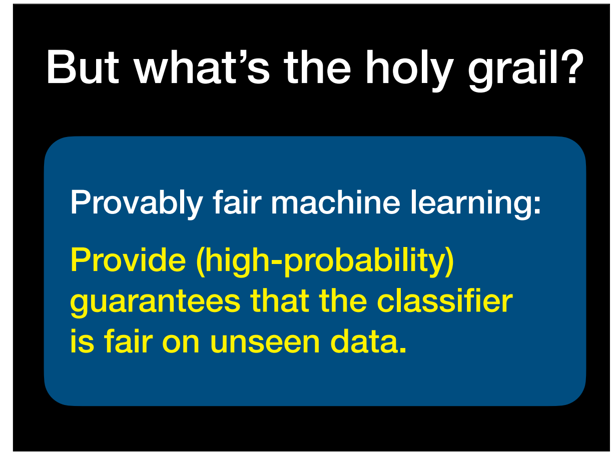
How do you flip the race of a photo?

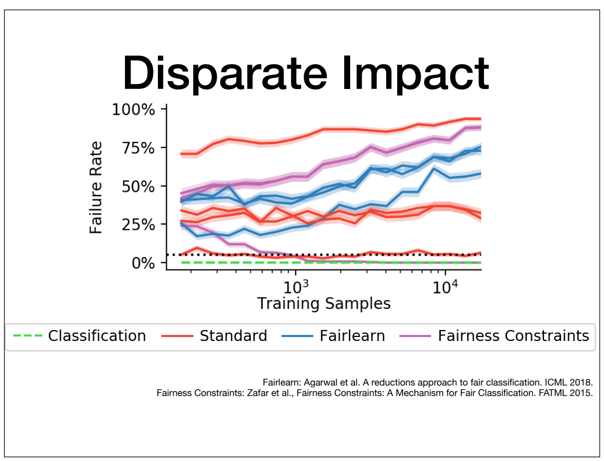


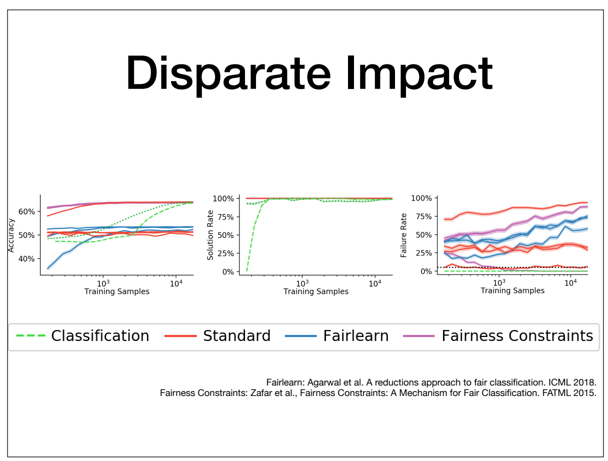
generative adversarial machine learning

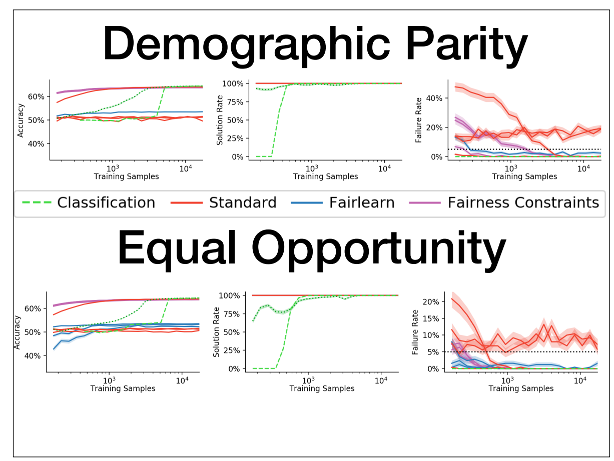

this-person-does-not-exist.com


What are we doing now?
Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots
Fair computer vision
He is a nurse.
She is a doctor.
O bir hemşire.
O bir doktor.
O bir hemşire.
O bir doktor.
She is a nurse.
He is a doctor.
Fair natural language processing
applied

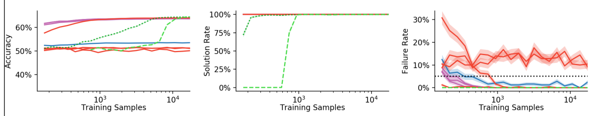

But what's the holy grail?
Provably fair machine learning:
Provide (high-probability) guarantees that the classifier is fair on unseen data.


Disparate Impact
Failure Rate
100%
75%
50%
25%
0%
10^3
10^4
Training Samples
Classification
Standard
Fairlearn
Fairness Constraints
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
Accuracy
Solution Rate
Failure Rate
Training Samples
Classification
Standard
Fairlearn
Fairness Constraints
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.

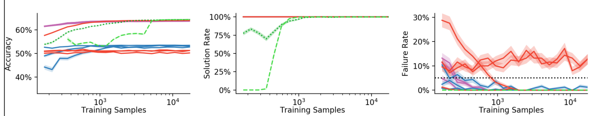

Demographic Parity
Accuracy
Solution Rate
Failure Rate
Training Samples
Classification
Standard
Fairlearn
Fairness Constraints
Equal Opportunity
Accuracy
Solution Rate
Failure Rate
Training Samples

Equalized Odds



Classification — Standard — Fairlearn — Fairness Constraints

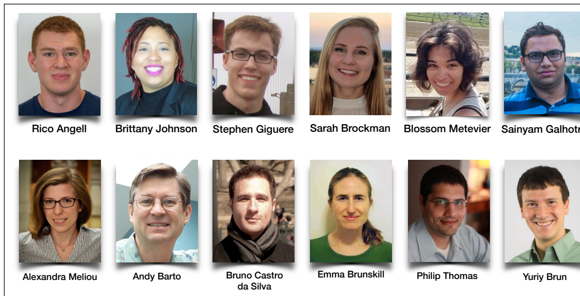
Predictive Equality



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 UMass Amherst 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