Bias in Software Systems

CS 520





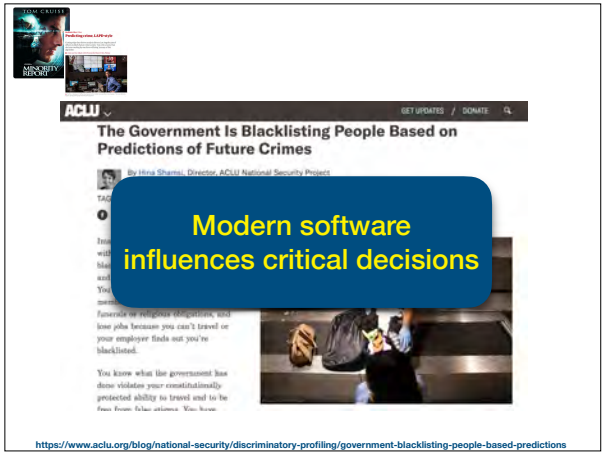
Resilient cities Cities

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

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



ACLU

The Government Is Blacklisting People Based on Predictions of Future Crimes

Modern software influences critical decisions

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

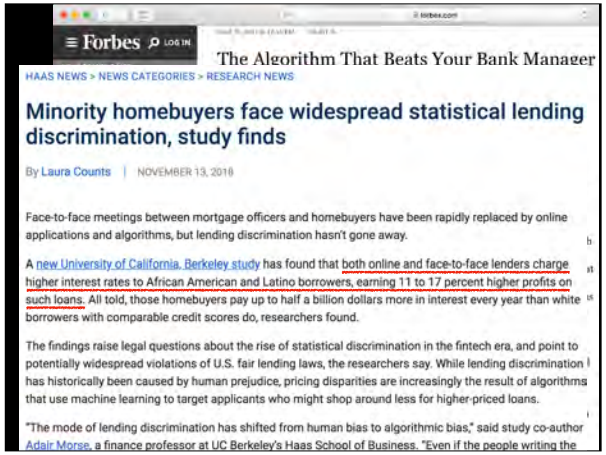


THE WALL STREET JOURNAL.

On Orbitz, Mac Users Steered to Pricier Hotels

On Orbitz, Mac Users See Costlier Hotel Options

THE CLOUD ALONE ISN'T "SMART."



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

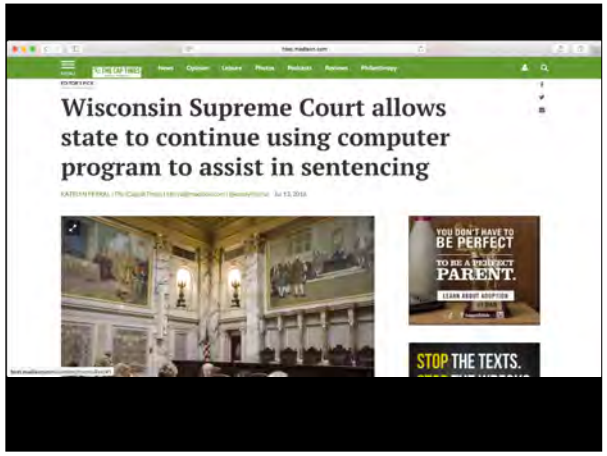
By Laura Counts | NOVEMBER 13, 2018

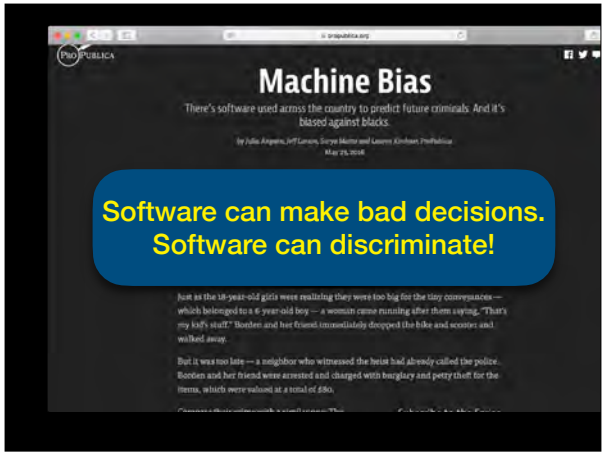
Face-to-face meetings between mortgage officers and homebuyers have been rapidly replaced by online applications and algorithms, but lending discrimination hasn't gone away.

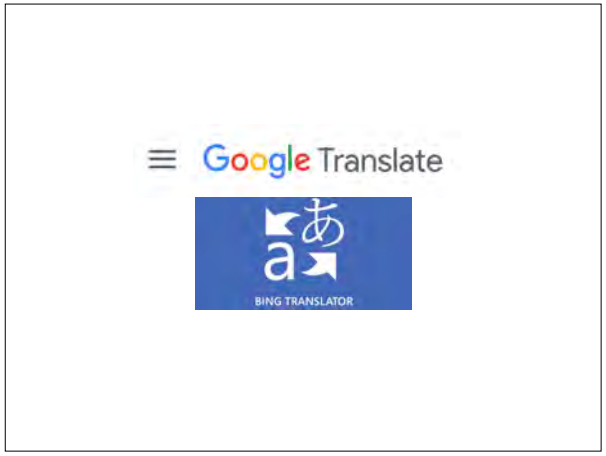
A new University of California, Berkeley study has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers, earning 11 to 17 percent higher profits on such loans. All told, those homebuyers pay up to half a billion dollars more in interest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

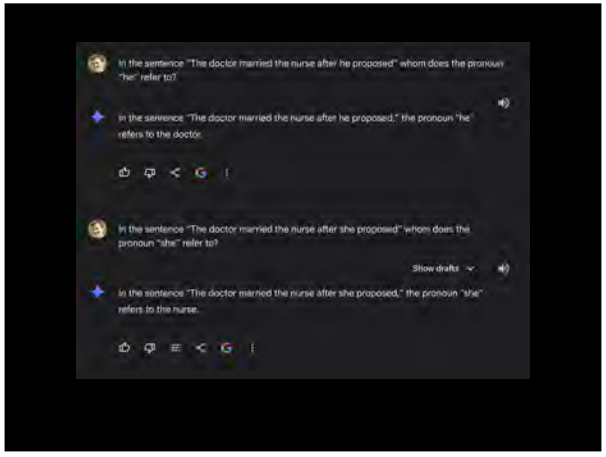
"The mode of lending discrimination has shifted from human bias to algorithmic bias," said study co-author Adair Morse, a finance professor at UC Berkeley's Haas School of Business. "Even if the people writing the

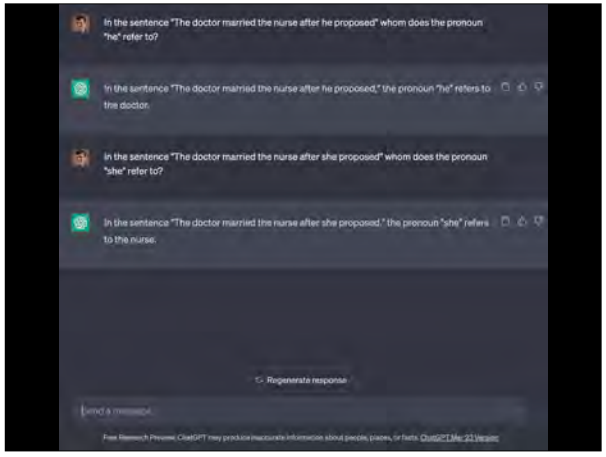

Wisconsin Supreme Court allows state to continue using computer program to assist in sentencing
YOU DON'T HAVE TO BE PERFECT TO BE A PERFECT PARENT.
STOP THE TEXTS.


ProPublica
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
Software can make bad decisions.
Software can discriminate!


Google Translate
BING TRANSLATOR


Microsoft Bing
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.

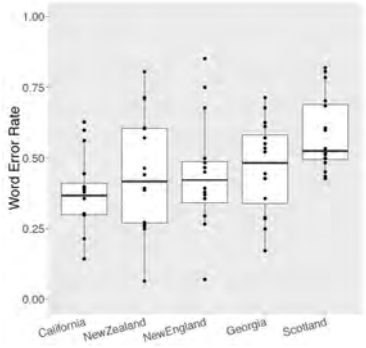

In the sentence "The doctor married the nurse after he proposed" whom does the pronoun "he" refer to?
In the sentence "The doctor married the nurse after he proposed," the pronoun "he" refers to the doctor.
In the sentence "The doctor married the nurse after she proposed" whom does the pronoun "she"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

YouTube

YouTube automatic captions

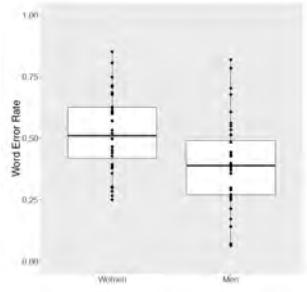


YouTube automatic captions



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

The Cost of Sounding Different: Accent Bias in Audio Language Models

"The AI models generally ranked Chinese and Nigerian accents the lowest, Indian in the middle, and American and British the highest. Even when the content was exactly the same and the models could accurately transcribe the speech, the systems ranked both the delivery and content higher for the American and British speakers."

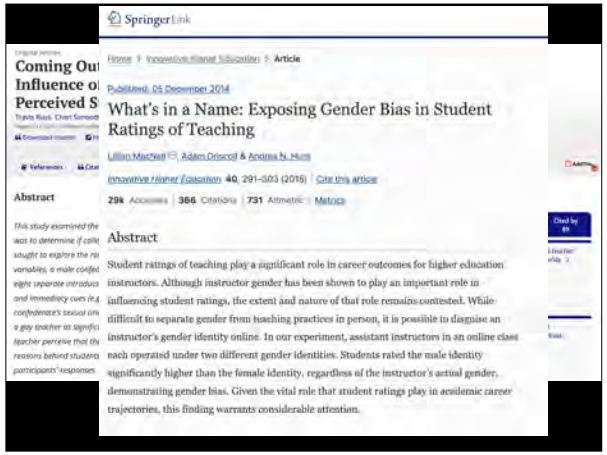


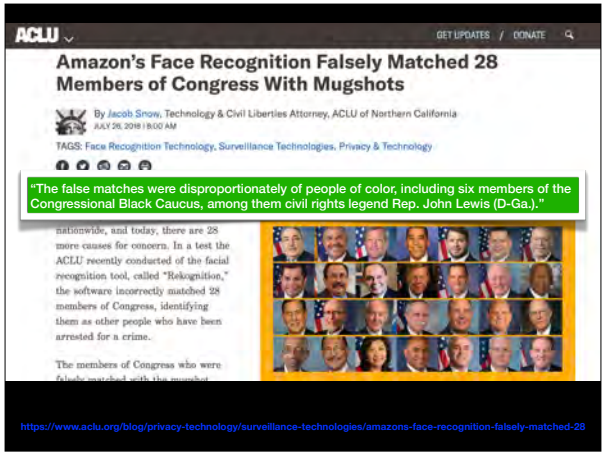
Neutrality Bites: Gender Representation in AI-Generated Animal Stories

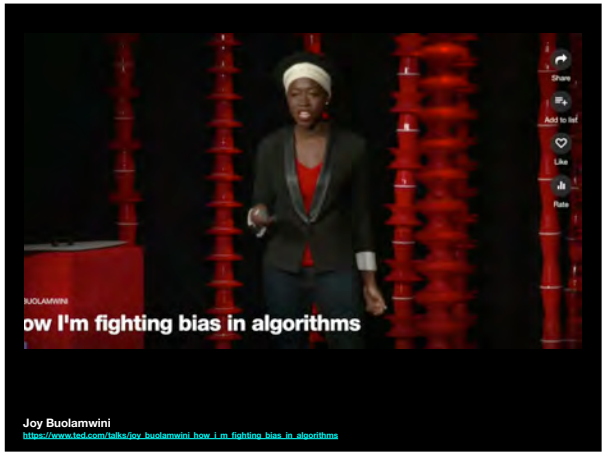
IMANI FINKLEY, University of Washington, USA
YUANXI LI, University of Washington, USA
MELANIE WALSH, University of Washington, USA

"The researchers gave six leading AI models variations on the same prompt they gave human participants. Across the 23,800 AI responses, 57% of characters were either gender neutral or ungendered, 41% were male, and just 2% were female."









how people want to use vision software



how people want to use vision software

HireVue

Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval

Kyra Wilson, Aylin Caliskan
University of Washington Information School
Seattle, Washington, USA
{kywi, aylin}@uw.edu

"The researchers varied names associated with white and Black men and women across over 550 real-world resumes and found the LLMs favored white-associated names 85% of the time, female-associated names only 11% of the time, and never favored Black male-associated names over white male-associated names."

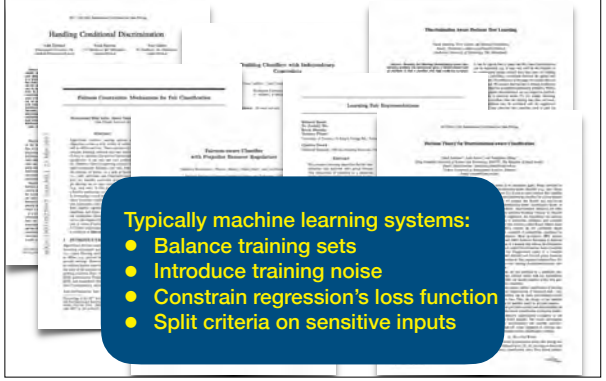
today's goals

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



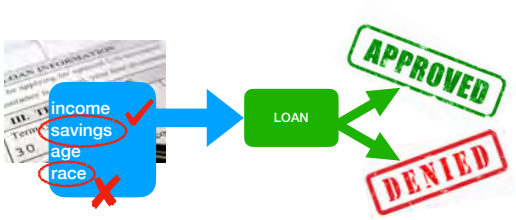
implementation bugs



unintended interactions and mismatched components

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

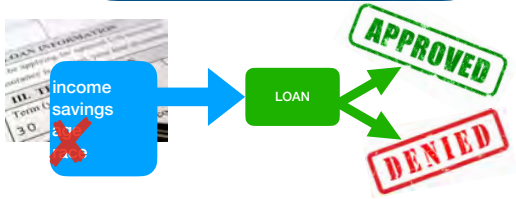
LOAN program



This talk is not about policy.

Fairness: Disparate Treatment

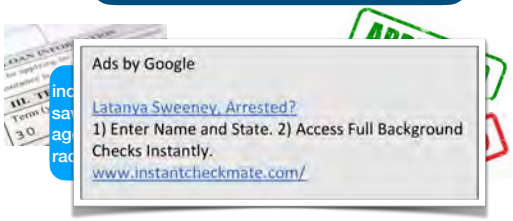
Hide the data



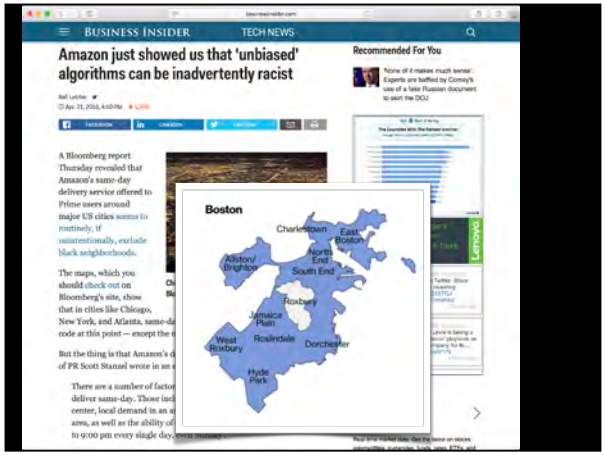
Zafar et al. Fairness constraints: Mechanisms for fair classification. AISTATS 2017.

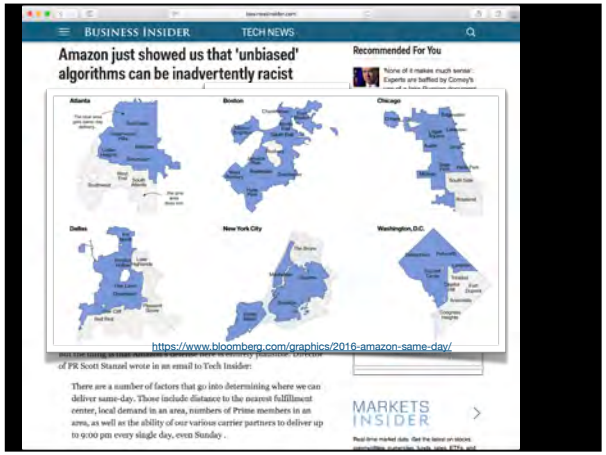
Fairness: Disparate Treatment

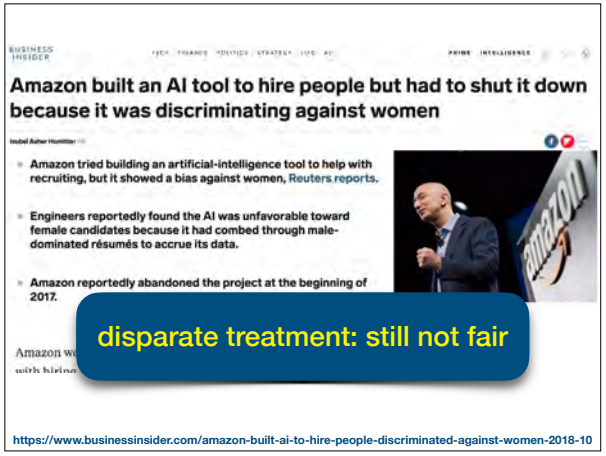
Hide the data



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







Fairness: Demographic Parity

Compare subpopulation proportions



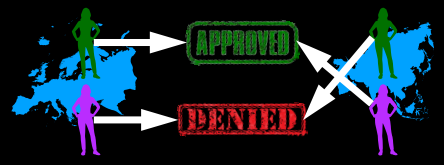
often called group discrimination

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

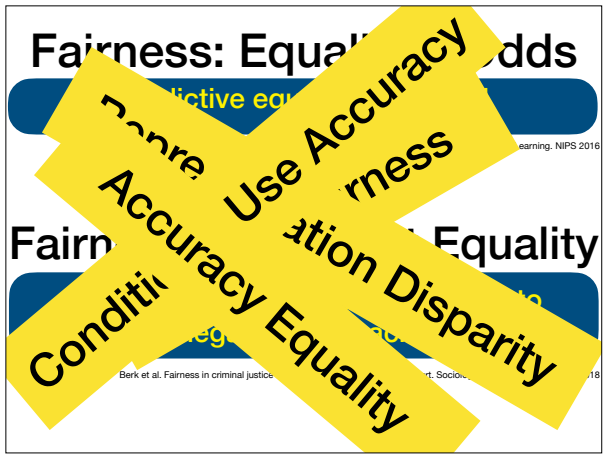
False positive rates should not differ

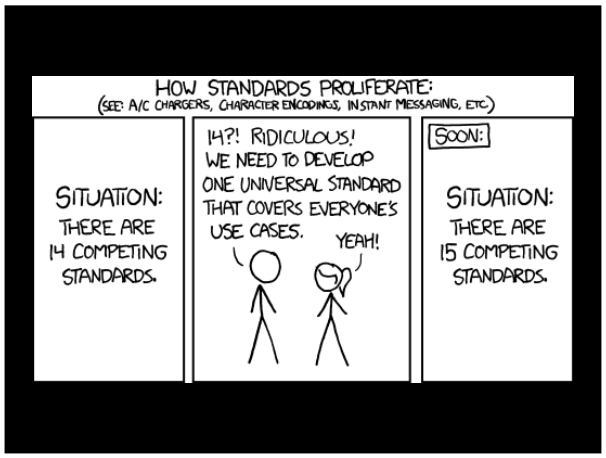
Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. FATML 2016
Corbett-Davies. Algorithmic decision making and the cost of fairness. KDD 2017

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: 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:



LOAN → ?

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

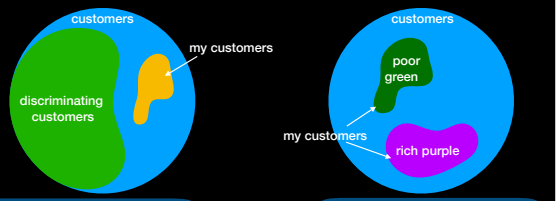
causal discrimination

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

But what's the holy grail?

Provably fair machine learning:

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

How machine learning works

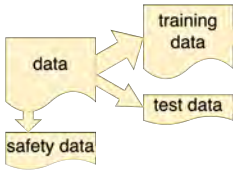


No guarantees on the behavior of the

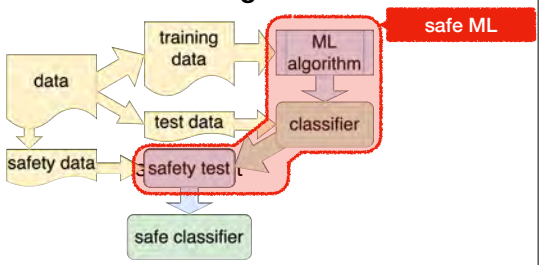
How machine learning should work



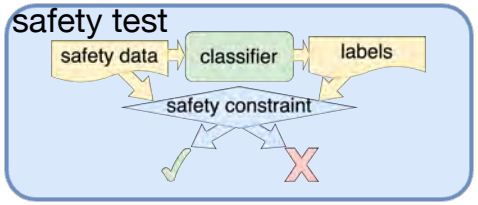
How machine learning should work



How machine learning should work



How machine learning should work

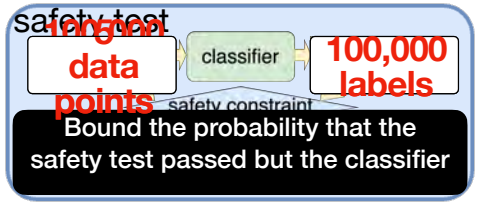


Example guarantees one might want

Use admission exam scores to predict success (GPA) at a university to guide admissions, but ensure that the

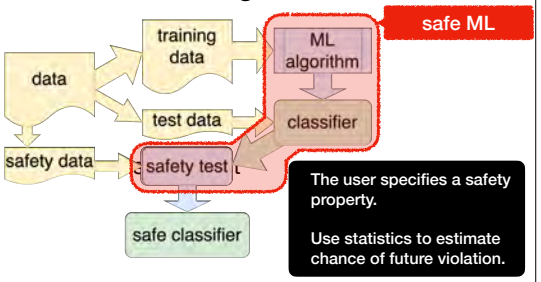
The frequency of false positive identifications in a facial recognition system is similar for people of every

How machine learning should work

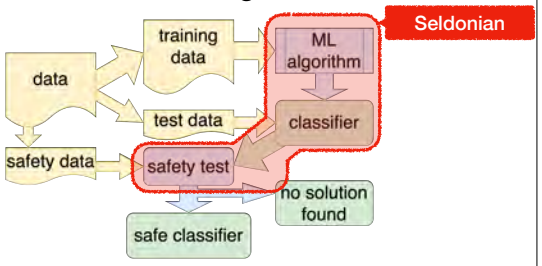


Bound the probability that the safety test passed but the classifier

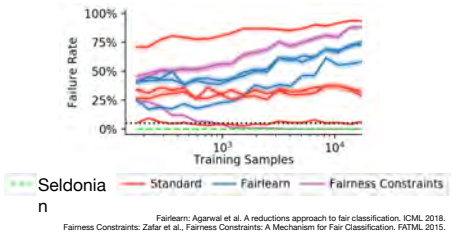
How machine learning should work



How machine learning should work

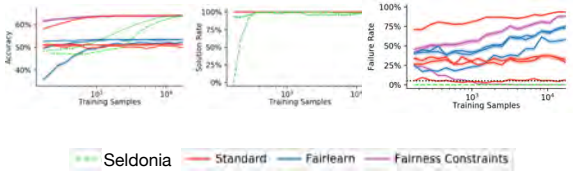


Seldonian ML enforcing fairness disparate impact



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.

Seldonian ML enforcing fairness 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.

Fairness



- 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