FUBON CENTER
AI IN BUSINESS SPEAKER SERIES

Machine Learning, Ethics, and Fairness

Dr. Solon Barocas in conversation with Prof. Foster Provost

WiFi Network: nyuguest
Wifi Login: sternguest | Password: gprokinf

#SternFubonAI

/NYUSternFubonCenter @NYUSternFC @nyustern

Machine Learning, Ethics, and Fairness

Solon Barocas

Microsoft Research and Cornell University

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

A new wave of start-ups — including Gild, Entelo, Textio, Doxa and GapJumpers — is trying various ways to automate hiring. They say that software can do the job more effectively and efficiently than people can. Many people are beginning to buy into the idea. Established headhunting firms like Korn Ferry are incorporating algorithms into their work, too.

If they succeed, they say, hiring could become faster and less expensive, and their data could lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases.

“Every company vets its own way, by schools or companies on résumés,” said Sheeroy Desai, co-founder and chief executive of Gild, which makes software for the entire hiring process. “It can be predictive, but the problem is it is biased. They’re dismissing tons and tons of qualified people.”

Miller (2015a)

"[H]iring could become faster and less expensive, and [...] lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods with high-interest loans.

Research from the University of Washington found that a Google Images search for "C.E.O." produced 11 percent women, even though 27 percent of United States chief executives are women. (On a recent search, the first picture of a woman to appear, on the second page, was the C.E.O. Barbie doll.) Image search results determined 7 percent of viewers' subsequent opinions about how many men or women worked in a field, it found.

Miller (2015b)

"But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result [...] algorithms can reinforce human prejudices."

Fairness, Accountability, and Transparency in Machine Learning

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."

The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

116TH CONGRESS
1ST SESSION

S. _____

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

IN THE SENATE OF THE UNITED STATES

Mr. WYDEN (for himself and Mr. BOOKER) introduced the following bill; which was read twice and referred to the Committee on _____

A BILL

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "Algorithmic Accountability Act of 2019".

SEC. 2. DEFINITIONS.

In this Act:

Bias as a technical matter

- Selection, sampling, reporting bias
- Bias of an estimator
- Inductive bias

- Of course, these raise ethical issues, too

Isn't discrimination the very point of machine learning?

- *Unjustified* basis for differentiation
- Practical irrelevance
- Moral irrelevance

Discrimination is not a general concept

- It is domain specific
  - Concerned with important opportunities that affect people's life chances
- It is feature specific
  - Concerned with socially salient qualities that have served as the basis for unjustified and systematically adverse treatment in the past

Regulated domains

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)
- ‘Public Accommodation’ (Civil Rights Act of 1964)

- Extends to marketing and advertising; not limited to final decision
- This list sets aside complex web of laws that regulates the government

Legally recognized ‘protected classes’

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967);**Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Discrimination law: two doctrines

Disparate Treatment

Formal

or

Intentional

Disparate Impact

Unjustified

or

Avoidable

Discrimination law: two doctrines

Disparate Treatment

Formal

or

Intentional

Equality of opportunity

Disparate Impact

Unjustified

or

Avoidable

Minimized inequality of outcome

Accounting for the optimism

The incidence and persistence of discrimination

- Callback rate 50% higher for applicants with white names than equally qualified applicants with black names
  - Bertrand, Mullainathan (2004)
- No change in the degree of discrimination experienced by black job applicants over the past 25 years
  - Quillian, Pager, Hexel, Midtbøen (2017)

- Formal procedures can limit opportunities to exercise prejudicial discretion or fall victim to implicit bias

Machine learning as the pinnacle of formal decision-making?

- Only what the data supports?
- Withhold protected features?
- Automate decision-making, thereby limiting discretion?

How *machines* learn to discriminate

How *machines* learn to discriminate

Skewed sample

Tainted examples

Limited features

Sample size disparity

Proxies

Barocas, Selbst (2016)

Skewed sample

- Police records measure “some complex interaction between criminality, policing strategy, and community-policing relations”
  - Lum, Isaac (2016)

Skewed sample: feedback loop

- Future observations of crime confirm predictions
- Fewer opportunities to observe crime that contradicts predictions
- Initial bias may compound over time

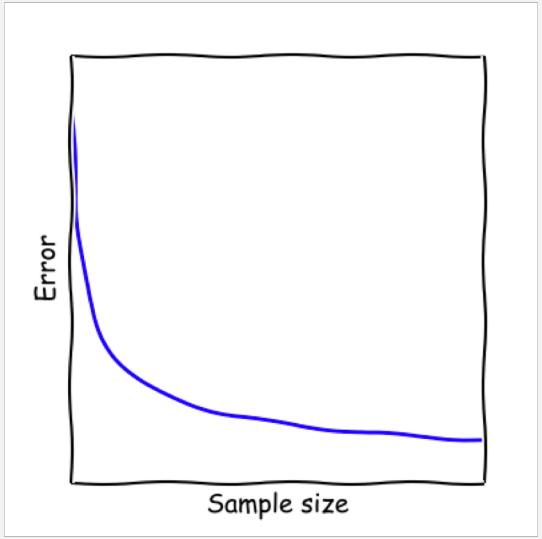
Tainted examples

- Learn to predict hiring decisions
- Learn to predict who will succeed on the job (e.g., annual review score)
- Learn to predict how employees will score on objective measure (e.g., sales)

Limited features

- Features may be less informative or less reliably collected for certain parts of the population
- A feature set that supports accurate predictions for the majority group may not for a minority group
- Different models with the same reported accuracy can have a very different distribution of error across population

Sample size disparity
Hardt (2014)



Proxies

- In many cases, making accurate predictions will mean considering features that are correlated with class membership
- With sufficiently rich data, class memberships will be unavoidably encoded across other features

How *machines* learn to discriminate

Skewed sample

Tainted examples

Limited features

Sample size disparity

Proxies

Barocas and Selbst (2016)

Three different problems

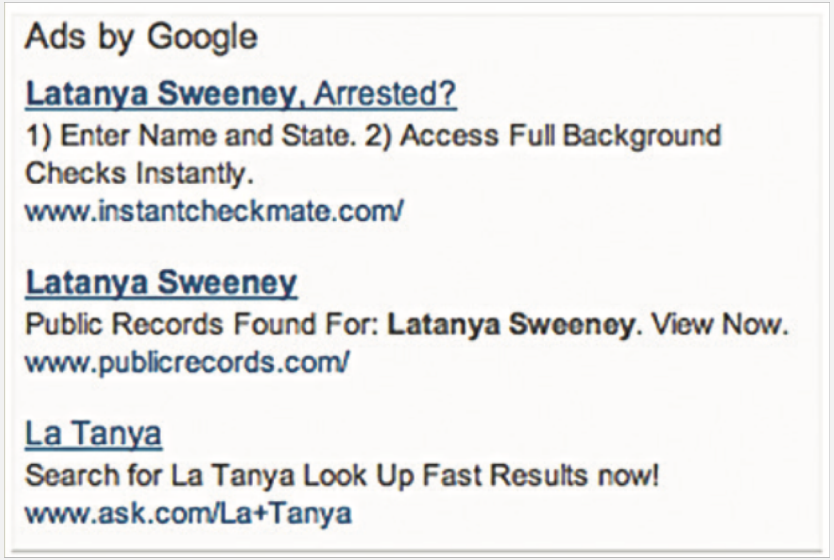
Discovering unobserved differences in performance
Skewed sample
Tainted examples

Coping with observed differences in performance
Limited features
Sample size disparity

Understanding the causes of disparities in predicted outcome
Proxies

From allocation to representation

Sweeney (2013)



Allocation

Criminal Justice

Employment

Credit

Housing

...

Representation

Representations of black criminality

Racial stereotype

Prospects in the labor market

Representation

Representations of black criminality

Racial stereotype

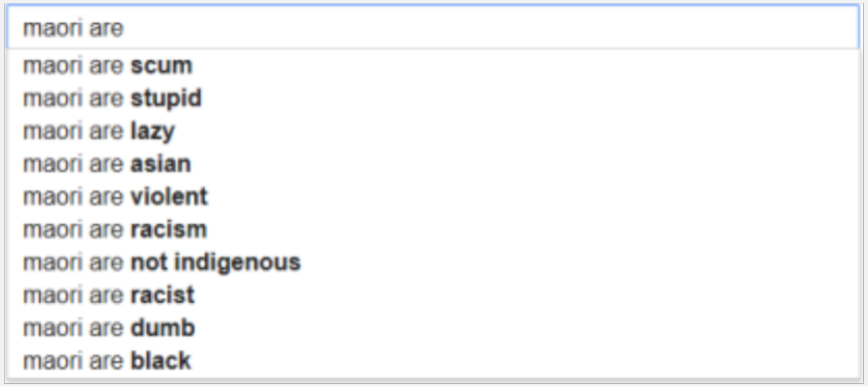
The problem with bias

Allocation	**Representation**
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

Buolamwini and Gebru (2018)

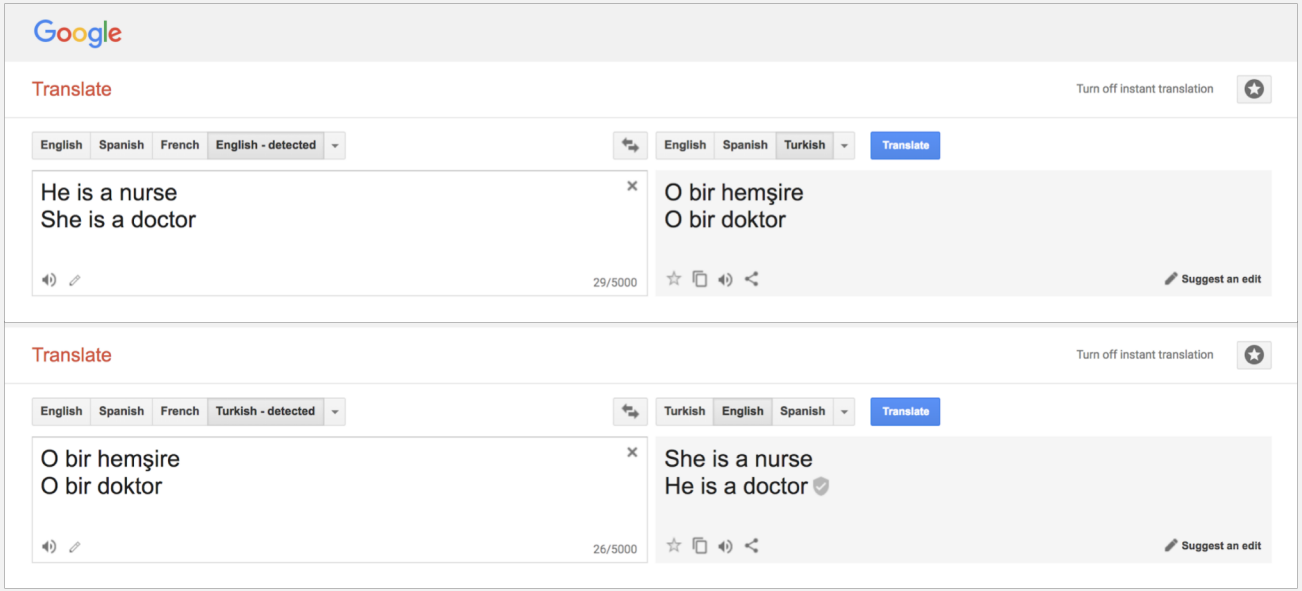
Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Elers (2014)


maori are
maori are scum
maori are stupid
maori are lazy
maori are asian
maori are violent
maori are racism
maori are not indigenous
maori are racist
maori are dumb
maori are black

Kay, Matuszek, Munson (2015)

Caliskan, Bryson, Narayanan (2017)



Potential responses

Do nothing

Reflects reality; if anything, exposes bias

You should know about the persistence of bias in the world and do something about it. Intervening denies people the opportunity to know about it and therefore do anything about it.

Improve accuracy

More data; higher quality data; better learning methods

Bias is an instance of error. You should solve it by continuing to improve accuracy.

Blacklist

Manually remove problematic labels; rely on users to identify troubling cases

You can solve problems by crowdsourcing the identification of bias and responding to it manually.

Scrub to neutral

Attempt to break or remove the problematic associations

You can achieve neutrality by severing biased associations; we can devise a method for preserving acceptable associations and removing unacceptable ones.

Representativeness

Bring in line with current reality; representations that are statistically representative

You should make representations adhere to true distributions in society.

Equal representation

Represent groups in equal proportion

You should present distributions as you would like them to be.

Awareness

Warn users that representations might reflect unfortunate dynamics in society

Humans will always have implicit biases but these can be overcome by making them aware. You should apply “warning labels” that prompt reflection on the implicit bias of algorithms.

FUBON CENTER
AI IN BUSINESS SPEAKER SERIES

Machine Learning, Ethics, and Fairness

Dr. Solon Barocas in conversation with Prof. Foster Provost

WiFi Network: nyuguest
Wifi Login: sternguest | Password: gprokinf

#SternFubonAI

/NYUSternFubonCenter @NYUSternFC @nyustern