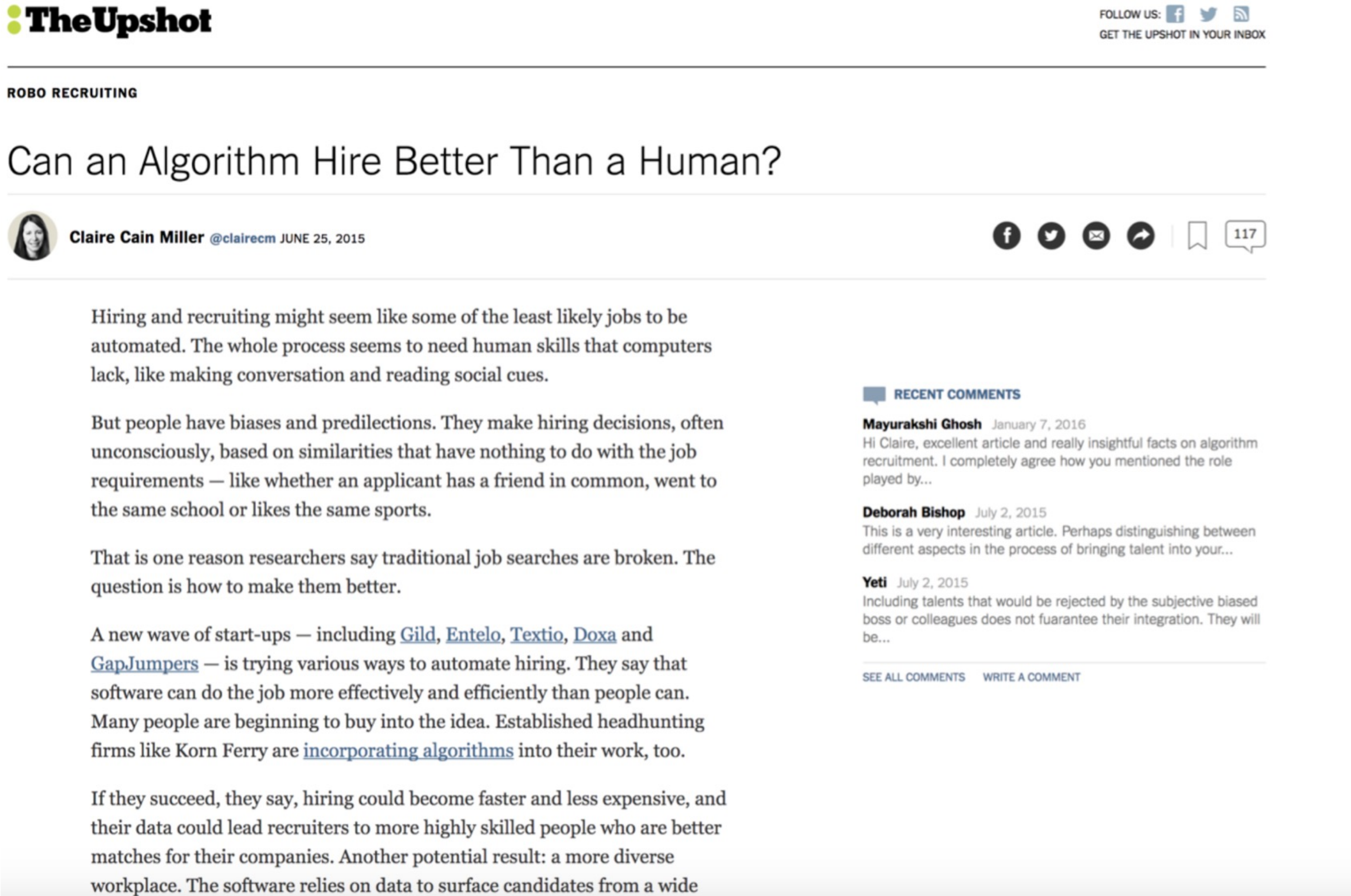
Logistics

- Submit peer reviews by 6pm.

- Milestone 2: Due Nov 5 (tomorrow)
  - Describe your progress in no more than 2 pages

Lecture 18
Fairness in AI

Instructor: Chien-Ju (CJ) Ho



TheUpshot

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.

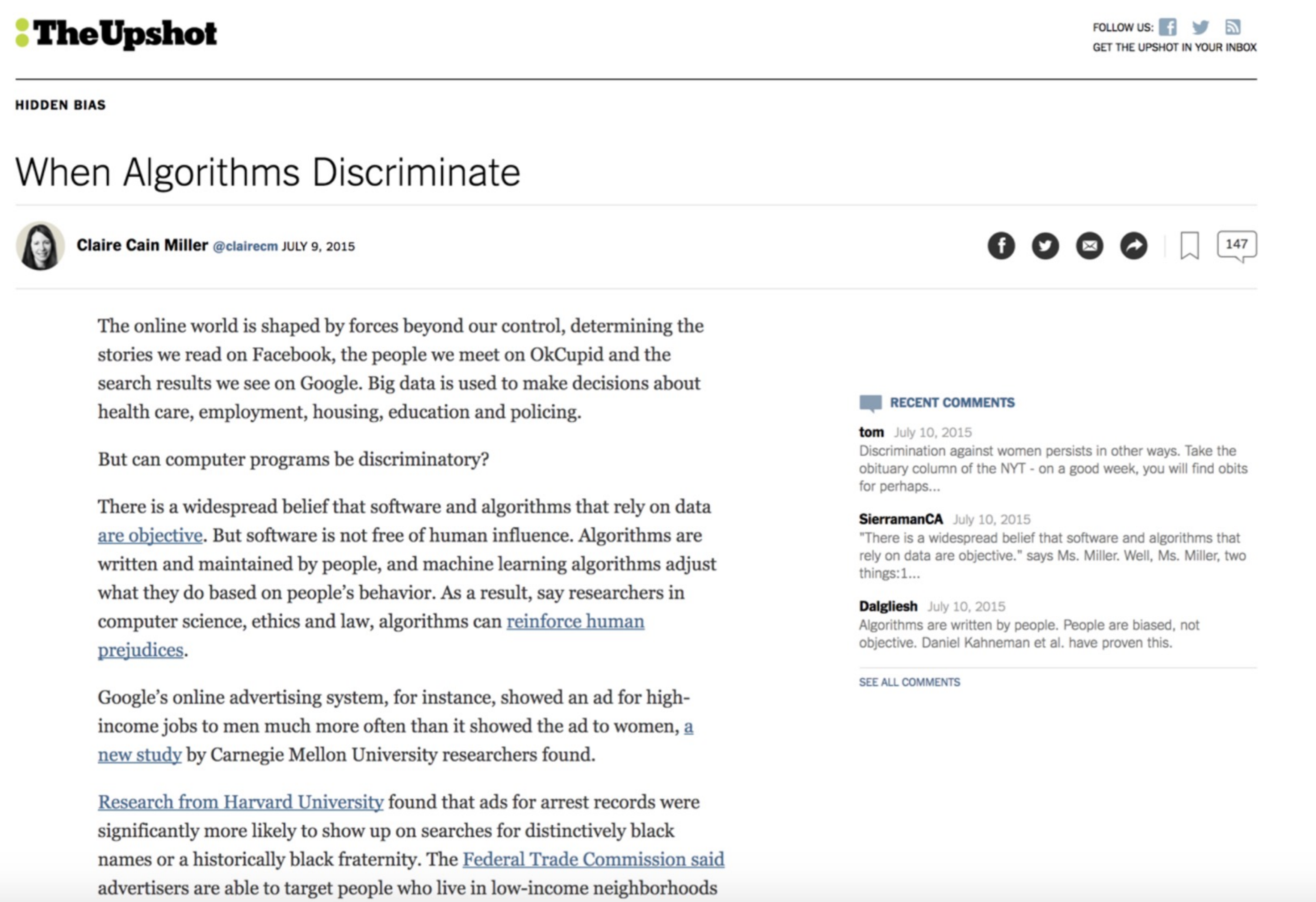
If they succeed, they say, hiring could become faster and less expensive, and their data could lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide

RECENT COMMENTS

Mayurakshi Ghosh January 7, 2016
Hi Claire, excellent article and really insightful facts on algorithm recruitment. I completely agree how you mentioned the role played by...

Deborah Bishop July 2, 2015
This is a very interesting article. Perhaps distinguishing between different aspects in the process of bringing talent into your...

Yeti July 2, 2015
Including talents that would be rejected by the subjective biased boss or colleagues does not fuarantee their integration. They will be...



TheUpshot

FOLLOW US:

GET THE UPSHOT IN YOUR INBOX

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

147

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

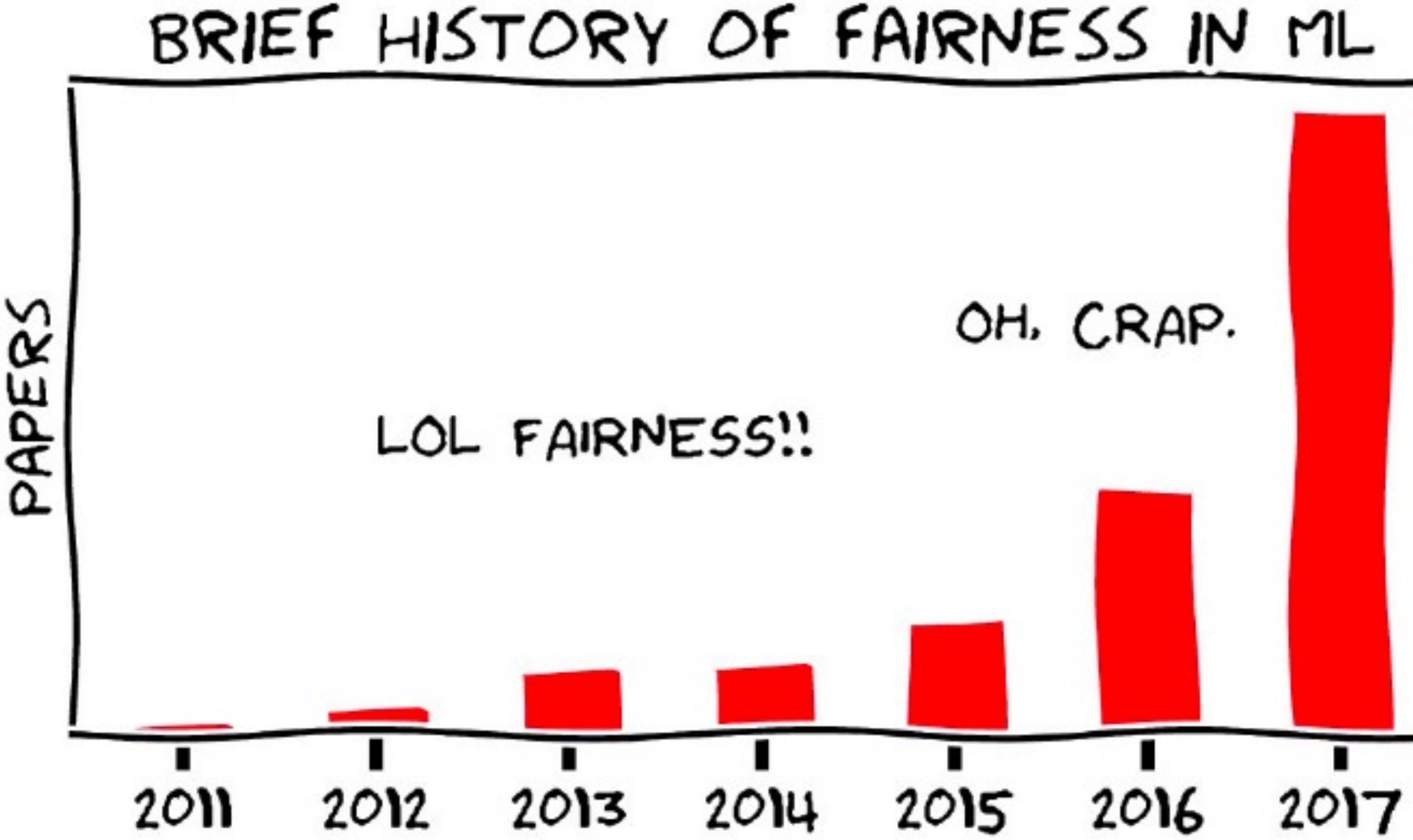
RECENT COMMENTS

tom July 10, 2015
Discrimination against women persists in other ways. Take the obituary column of the NYT - on a good week, you will find obits for perhaps...

SierramanCA July 10, 2015
"There is a widespread belief that software and algorithms that rely on data are objective." says Ms. Miller. Well, Ms. Miller, two things:1...

Dalgliesh July 10, 2015
Algorithms are written by people. People are biased, not objective. Daniel Kahneman et al. have proven this.

SEE ALL COMMENTS


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017

Isn’t the point of ML to discriminate?

Want to avoid “unjustified” discrimination

Example: Loan Applications

- By law, the banks can’t discriminate people according to their race.
- First natural approach (fairness through blindness)
  - remove the race attribute from the data
- Guess what happened?
  - Redlining

BANK
ATM
NO
Mulberry Baptist Church

What should we do?

- From computer scientists / engineers’ point of view….
- Give me an operational definition of fairness, I’ll implement a system that satisfy it!

- How should we define fairness?
- Is it even possible to define a universal fairness notion?

This Lecture: Group Fairness

- Consider social groups defined by sensitive attributes, such as gender, race, sexual orientation, etc.

- If we believe a decision shouldn’t be depending on these sensitive attributes, then we want the outcome for different groups to be similar.

Won’t Recidivate	**TN1**	**FP1**
Will Recidivate	FN1	TP1
	Labeled Low-Risk	Labeled High-Risk

Won’t Recidivate	**TN2**	**FP2**
Will Recidivate	FN2	TP2
	Labeled Low-Risk	Labeled High-Risk

- Defendant: the probability that I’m incorrectly classified high-risk is independent of my race.
  - Equal False Positive Rate: $\frac{\mathrm{FP1}}{\mathrm{FP1}+\mathrm{TN1}} = \frac{\mathrm{FP2}}{\mathrm{FP2}+\mathrm{TN2}}$
- Defendant: the probability that I’m incorrectly classified as low-risk is independent of my race.
  - Equal False Negative Rate: $\frac{\mathrm{FN1}}{\mathrm{FN1}+\mathrm{TP1}} = \frac{\mathrm{FN2}}{\mathrm{FN2}+\mathrm{TP2}}$
- Decision-maker: the ratio of people who recidivated among the ones labeled high-risk is independent of race.
  - Equal predictive value: $\frac{\mathrm{TP1}}{\mathrm{TP1}+\mathrm{FP1}} = \frac{\mathrm{TP2}}{\mathrm{TP2}+\mathrm{FP2}}$

Impossibility Result [Kleignberg et al. 2016]

The above three conditions cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.

There are more definitions of fairness

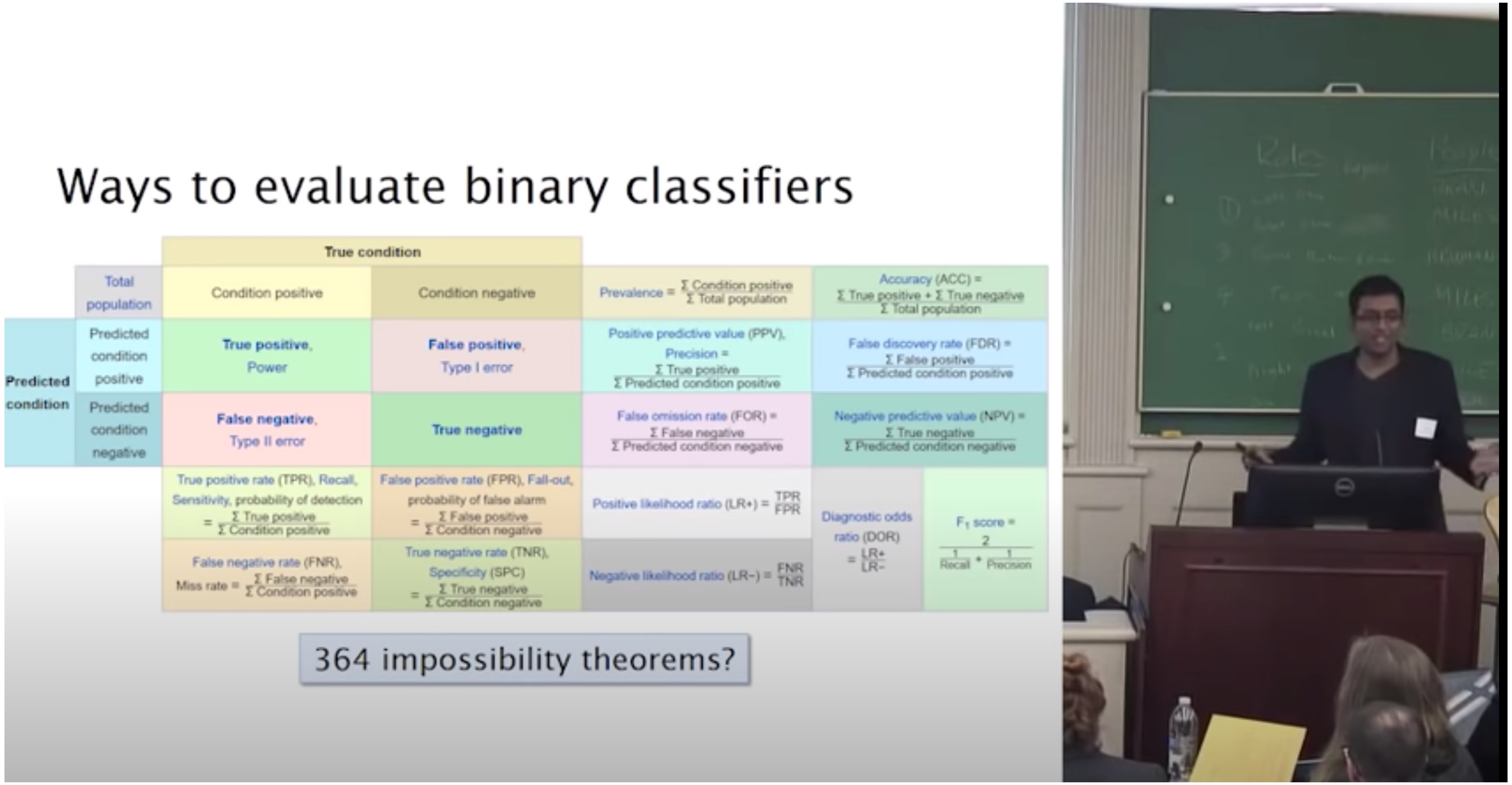
- If there are intrinsic differences between the groups, is it fair to ignore them?

- Another setup
  - A: Sensitive attributes (e.g., race)
  - Y: True labels (e.g., commit a crime in the future)
  - C: Predictions (e.g., predictions of recidivism)
- Criteria:
  - C independent of A
  - C independent of A conditional on Y
  - Y independent of A conditional on C

The same impossibility results exist!

There are more possible definitions of fairness

- FAT*18 Tutorial: 21 Definitions of Fairness and Their Politics. Arvind Narayanan.



Ways to evaluate binary classifiers

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

364 impossibility theorems?

Individual Fairness

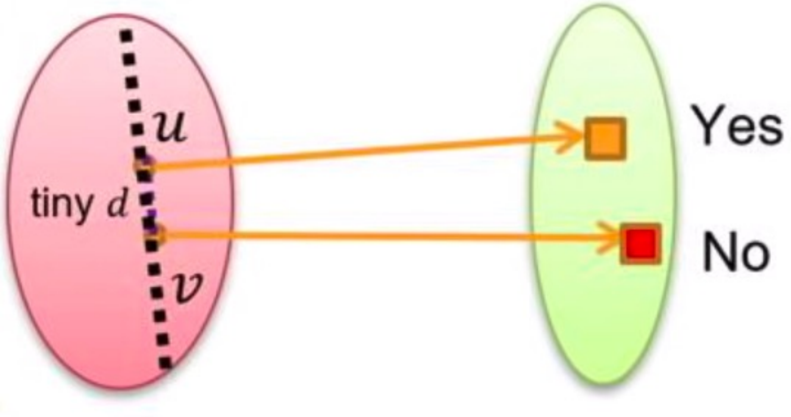
- Cucumbers and Grapes: https://youtu.be/-KSryJXDpZo

Individual Fairness

- Similar people should be treated similarly

- Challenges
  - What do we mean by similar people
    - Need to define some kind of “distance” measure

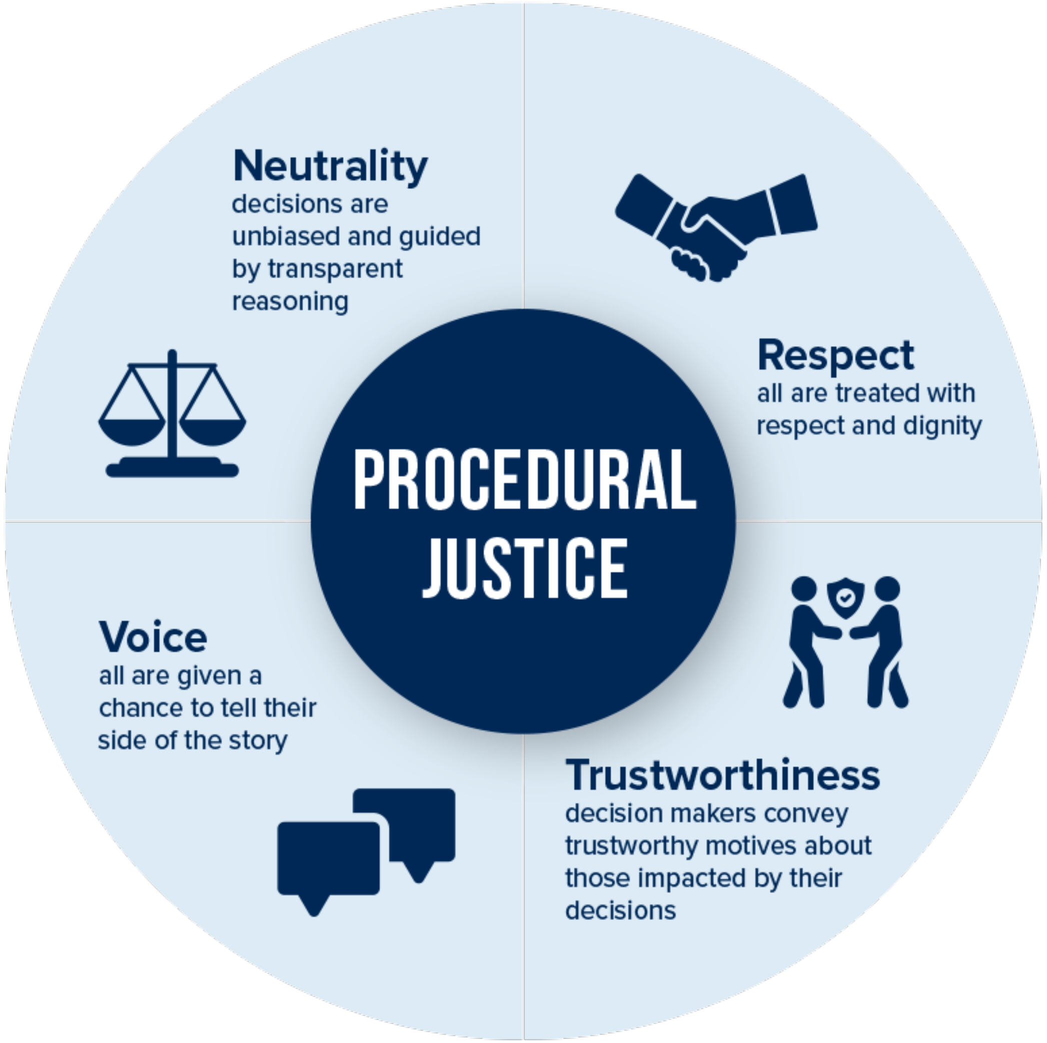
  - What do we mean by being treated similarly
    - Decisions based on threshold won’t work
    - Need to impose some “smooth” notion
    - Randomization is often required



Counterfactual Fairness

- A decision is fair towards an individual if it gives the same predictions in
  - (a) the observed world and
  - (b) a world where the individual had always belonged to a different demographic group

Procedural Fairness (Procedural Justice)



PROCEDURAL JUSTICE

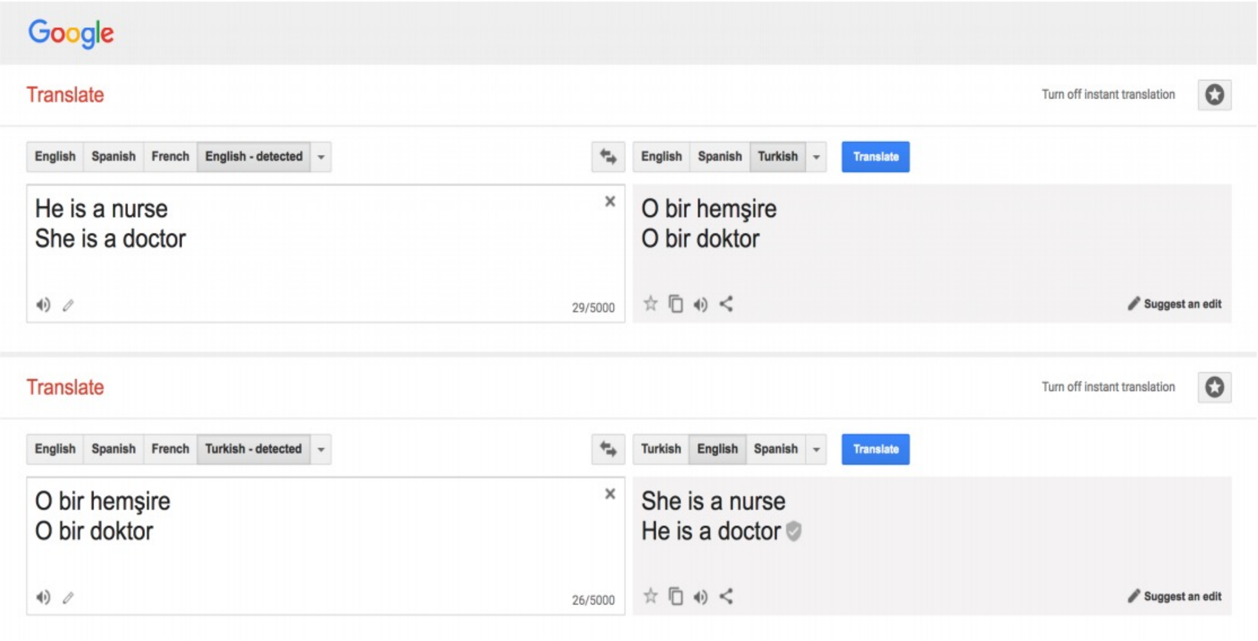
Neutrality
decisions are unbiased and guided by transparent reasoning

Respect
all are treated with respect and dignity

Voice
all are given a chance to tell their side of the story

Trustworthiness
decision makers convey trustworthy motives about those impacted by their decisions

Representation

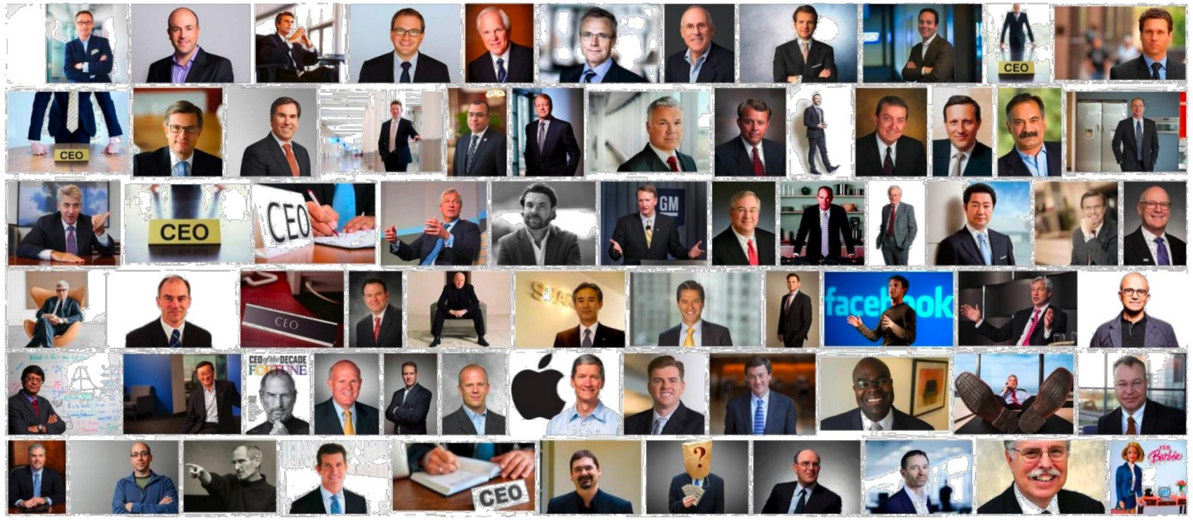


He is a nurse
She is a doctor

O bir hemşire
O bir doktor

O bir hemşire
O bir doktor

She is a nurse
He is a doctor



[Kay et al., 2015]

[Caliksan et al., 2017]

Take-Aways

- As AI/ML becomes more ubiquitous in our daily decision making, ethical issues are getting more important as well.
- **Being aware** of the issues is the important first step!
- ”Solving” the issues (if at all possible) requires communications among people in different disciplinaries.

- More references:
  - NIPS17 Tutorial: https://vimeo.com/248490141
  - FAT*18 Tutorial: https://www.youtube.com/watch?v=jIXIuYdnyyk
  - FAT*19 Putting fairness in practice: https://algorithmicbiasinpractice.wordpress.com