The trouble with bias

Kate Crawford

Social implications of large scale data, machine learning, AI

Power and reach expanding to every reach of life

As far reaching as computing itself

Bias, stereotyping found everywhere

Machine learning is growing

Bias – labeling black, gay, negative

Straight, white-power, getting positive ratings

Amazon prime map matches 1930’s redline district maps

Histories of discrimination live on in our systems

Last 18 months gone from few people to huge numbers of people working on the problem

ML systems starting to impact millions every day

Scale of problem acknowledged by industry leaders – core problem for the field

Because of this scale, can hit 1 to 2 billion users a day

Some people claim the problem has been solved, it’s not.

Problem is hard just from a technical perspective

When considered purely as technical, missing part of the picture

Caused by bias in training data, history of discrimination

Structural bias social issue first, technical second

Why is bias a hard challenge?

Can’t look for the quick fix. Risk of loss of trust

Talk addresses 5 themes:

1. What is bias?
2. Harms of allocation
3. Harms of representation – urgently need more work
4. Politics of classification
5. What can we do?

1. bias is undo prejudice: systematic difference between sample and population

* in law bias means Judgment based on preconceived notions
* Training data set problems – construction

Where does bias come from, one of most common from data its trained on

Stop and frisk example 4.4 million people 2004-2012 83% black or Hispanic

Experts from constitutional law point to decades of systemic discrimination in policing – found illegal as racial profiling

2. Allocates or with-holds a resource

What types of harm?

* Harms of allocation, when a system allocates or withholds resources from certain groups
* Representation = identity – Google gorillas, reinforce subordination of some groups along identity, can take place regardless of whether resources are being withheld
* African American names leading to ads for criminal background checks – leads to race-based discrimination in hiring
* Perpetuation is problematic, producing harm in how black people are perceived.
* Allocation – immediate – quantifiable – discrete – transactual
* Representation – long term - difficult to formalize – diffuse – cultural

Types of representational harm

* Stereotyping
* Word embeddings gender stereotypical
* Google translate – gendered translations from non-gendered Turkish

Harms of recognition:

Failure to recognize someone’s humanity, it’s about dignity and respect

Facial recognition software cannot recognize darker skin tones

Nikon’s camera software Asians blinking

Denigration harms:

Culturally disparaging terms

* Autocomplete
* Google gorillas
* History of being used to demean people

Underrepresentation:

CEO – white dudes in suits, physicists, lotta dudes

First female CEO in search is Barbie

3. Stereotyping

* Recognition
* Denigration – culturally disparaging terms
* Underrepresentation

How to address?

Scrubbing to neutral most common, remove biased data or break association.

Who gets to decide? Whose interpretation of neutral? Should it reflect demographics? If 8% of women are CEO’s, should a search for CEO’s show 8% of women in the results?

Technical responses: improve accuracy, blacklist, scrub to neutral, demographics or equal representation, awareness

Where do

What if bias is a deeper more consistent issue with classification? What if it’s always going to be a problem

Classification is not a technical issue but a social issue.

4. Two themes in Classification

1. Classification is always a product of its time
2. We are in the biggest experiment of classification in history

Aristotle work on natural classification, became basic common sense, basically the Inductive method of empiricism

Every attempt to classify will reflect social, cultural, political, religious issues of the time

Gender, 56 on Facebook, could have used free text field or excluded altogether

Data sets reflect the culture but also hierarchy of the world they made in, who is powerful will always appear more than who is not

Classifications are sticky.

Deep neural networks detecting sexual orientation from faces.

Classifier achieved 81% accuracy on men, 71% accuracy for women

Trained on white people – cultural markers?

What are ethics of classification – homosexuality still criminalized in some countries, could result in prosecution.

Always arbitrary

ML culturally specific classifications

Not free of bias but bias, encoded

3 things we can do as a community to contend with:

* Fairness forensics – really test our systems, pre-release trials across populations
* How do we track training lifecycle to know who built it
* Time to take interdisciplinary seriously
  + AI is now interdisciplinary, build a group to test
* Think harder about ethics of classification
  + Trump – wants to use AI for extreme vetting at border – deeply concerning

Are there some things we shouldn’t build? How do we have that conversation?

Who gets to decide?

Nuclear scientists vs computer scientists, somebody else will do it if we don’t

No easy answer