## HIM Lecture 3 What Is Machine Bias? Can and Should It Be Avoided?

Joanna J. Bryson University of Bath, United Kingdom

@j2bryson

Many people are in awe of AI and robots. They expect machines will be perfect, unbiased, immortal, omniscient.



Robots that are made with sufficient precision and to be sufficiently stupid can do repetitive tasks remarkably quickly and accurately.

That's nothing like omniscience.

BMW plant in Leipzig, Germany: Spot welding of BMW 3 series car bodies with KUKA industrial robots. via Wikimedia

Al Is Not Magic Deep Learning Is Not Magic No Learning is Magic

Computation is a physical process, not an abstraction like math. It takes time, space, and energy

## Combinatorics and Tractability

- There are more possible short chess games than atoms in the universe.
- Biology has a lot more options than chess.
- Human uniqueness derives from our unique (in extent) capacity to pool the outcomes of our computation.



From: The Thomas Crown Affair

## One Consequence Al Is Not Necessarily Better than We Are



Semantics derived automatically from language corpora contain human-like biases Aylin Caliskan, Joanna J. Bryson and Arvind Narayanan (April 13, 2017) Science 356 (6334), 183-186. [doi: 10.1126/science.aal4230]

# Semantics: What Does meaning Mean?

How can we know what words mean? Hypothesis: a word's meaning is no more or less than how it is used.



(Quine 1969)

#### From the 1990s

## Large Corpus Semantics

- We can learn how a word is used (its meaning, or semantics) by parsing normal language (Finch 1993, Landauer & Dumais 1997, McDonald & Lowe 1998).
- Record co-occurring words (those nearby on either side of the target word).
- Store counts for 75 fairly frequent words...
  - $\Rightarrow$  'Meaning' is cosine in 75-D space.



#### **OLD WAY**

Cosines between semantic vectors correlate with human reaction times (Figure: 75-D space projected in to 2-D, McDonald & Lowe 1998)

salt	circle	gold	month	measles
	SI	ver		
	square			
			year	
lightning	3			
		stor		eister
	abba	star vi l	atin	sister
	cabbag	je i	aun	
lettuce				
lettace	dag			
	0.5	greek		aueen
	cat	Breen		dagen
			soldier	
thun	der			
black	moon	sailo	r	
white	m	umps	king	brother

## Implicit Association Task

Greenwald, McGhee, & Schwartz (1998) cf. Bilovich & Bryson (2008), Macfarlane (2013)

NEW WAY



Associated concepts are easier to pair Differential reaction time is a measure of bias Slides with these fonts courtesy Arvind Narayanan

## Hypothesis: corpus semantics will capture these same biases



distance between means is measured in standard deviations (d)

Report:

1.effect size measured in d (known to be huge for human IAT) 2.probability of sets of terms being same population (p value) Hypotheses: corpus semantics will capture these same biases

Al Built with ML Contains Our Implicit Biases Implicit Biases Are a Part of Ordinary Semantics

## Corpus, training, and stimuli all established standards

#### Common crawl: web corpus

- -840 billion tokens
- -2.2M unique

All "off the shelf" Exploring standard effects in existing, widely-used AI tools

GloVe

- -Stanford project, state of the art
- Pre-trained embeddings
- 300-dimensional vectors

[Very similar results with word2vec/Google News]

### FINDINGS

#### Warmup: universal biases Greenwald, McGhee, & Schwartz (1998)

Flowers: aster, clover, hyacinth, marigold	Insects: ant, caterpillar, flea, locust
Pleasant: caress, freedom, health, love	Unpleasant: abuse, crash, filth, murder

Original finding [N=32 participants]:d = 1.35,  $p < 10^{-8}$ Our finding [N=25x2 words]:d = 1.50,  $p < 10^{-7}$ 

## Racial bias [valence]

Greenwald, McGhee, & Schwartz (1998)

European-American names: Adam, Harry, Josh, Roger, ... African-American names: Alonzo, Jamel, Theo, Alphonse...

Pleasant: caress,Unpleasant: abuse, crash,freedom, health, love...filth, murder...Original finding [N=26 participants]:d = 1.17,  $p < 10^{-6}$ Our finding [N=32x2 words]:d = 1.41,  $p < 10^{-8}$ Our finding on the Bertrand & Mullainathan (2004) Résumé Study(assuming less pleasant  $\Rightarrow$  fewer invites):d = 1.50,  $p < 10^{-4}$ 

#### Gender bias [stereotype] Nosek, Banaji, & Greenwald (2002)

Female names: Amy, Joan, Lisa, Sarah... Male names: John, Paul, Mike, Kevin...

Family words: home, parents, children, family...

Career words: corporation, salary, office, business, ...

Original finding [N=28k participants]: d = 1.17,  $p < 10^{-2}$ Our finding [N=8x2 words]: d = 0.82,  $p < 10^{-2}$ 

#### Gender bias [stereotype] Nosek, Banaji, & Greenwald (2002b)

Science words: science, technology, physics, ... Arts words: poetry, arts, Shakespeare, dance...

Male words: brother, father, uncle, grandfather... Female words: sister, mother, aunt, grandmother ...

Original finding [N=83 participants]: $d = 1.47, p < 10^{-24}$ Our finding [N=8x2 words]: $d = 1.24, p < 10^{-2}$ 

Observe: Machine Learning can mine visceral "facts" about human qualia (e. q. insects are unpleasant) without direct experience of the world. The same process mines truth.

## Biases In the Web Can Be Accurate





**Figure 2.** People with androgynous names Pearson's correlation coefficient  $\rho = 0.84$  with *p*-value  $< 10^{-13}$ .

Basic Definitions Caliskan, Bryson & Narayanan 2017

- Bias: expectations derived from experience regularities in the world.
- Stereotype: biases based on regularities we do not wish to persist.
- Prejudice: acting on stereotypes.



#### Caliskan, Bryson & Narayanan 2017

- Bias: expectations derived from experienced regularities. Knowing what *programmer* means, including that most are male.
- Stereotype: biases based on regularities we do not wish to persist. Knowing that most programmers are male.
- Prejudice: acting on stereotypes. Hiring only male programmers.

## Critical Implication

- Bias: expectations derived from experience regularities in the world.
- Stereotype: biases based on regularities we do not wish to persist.
- Prejudice: acting on stereotypes.
- Stereotypes are culturally determined. No algorithmic way to discriminate stereotype from bias!

## At Least Three Sources of Al Bias

- Implicit: Absorbed automatically by ML from ordinary culture.
- Accidental: Introduced through ignorance by insufficiently diverse development teams.
- Deliberate: Introduced intentionally as a part of the development process (planning or implementation.)

### Does the source of bias matter?



- When a cheap plastic mask works better with a face detector than your actual face? (Joy Buolamwini, pictured)
- When searches for your name offer advertisements for criminal records at the top of the page? (Latanya Sweeney)
- When google images labels someone who looks vaguely like you as "gorilla"?
- When you can't use a soap dispenser without holding toilet paper in your hand?
- When you don't get parole?

## At Least Three Sources of Al Bias

- Implicit: Absorbed automatically by ML from ordinary culture.
- Accidental: Introduced through ignorance by insufficiently diverse development teams.
- Deliberate: Introduced intentionally as a part of the development process (planning or implementation.)

One Difference: How We Deal With Them

- Implicit—compensate with design, architecture (see also accidental).
- Accidental-diversify work force, test, log, iterate, improve.
  - Al really can be better than us sometimes, in that we can catch the mistakes and repair them, or limit data based on imperfect worlds.
- **Deliberate**—audits regulation, accountability.

HIM Lecture 7

## How should we address machine implicit bias?

Like we do our own.

## Architecting for Good

- Implicit Knowledge is statistics aggregated over a great number of examples / experiences (e.g. deep & reinforcement learning, latent semantic analysis.)
- Explicit Knowledge can be learned from one or a few presentations (relies on indexing into implicit knowledge, heuristic systems such as nearest neighbour, productions).
  - Associated with deliberate control.
  - Allows negotiation and rapid progress.
  - Can be programmed as rules (e.g. swear words in predictive text), cognisant error detection / self censorship, selective learning / inputs.

## It's in the Vectors; Can't We Just Fix It?

- Caliskan, Narayanan, & Bryson (2017): use a systems engineering approach that allows you to compensate for prejudice before acting.
- Bolukbasi, Chang, Zou, Saligrama, and Kalai (NIPS 2016): warp basic representation of semantics to conform to crowdsourced human expectations.
  - Such approaches assume biases are enumerable, and fairness desiderata are consistent and coherent. Neither is true.
- Fairness and ethics are a form of human cooperation an ever-changing (hopefully improving) complex negotiation of inconsistent human desires.

### Summary & Future

Artificial and Natural Intelligence are continuous with each other. Neutral Magic Færies of Mathematical Purity will not fix our problems.

Next lecture: AI  $\stackrel{?}{=}$  Humans

Next week!



### Thanks (coauthors on the Science paper)









Aylin Caliskan @aylin\_cim

Arvind Narayanan @random\_walker