There Is No AI Ethics

The Human Origins of Machine Prejudice

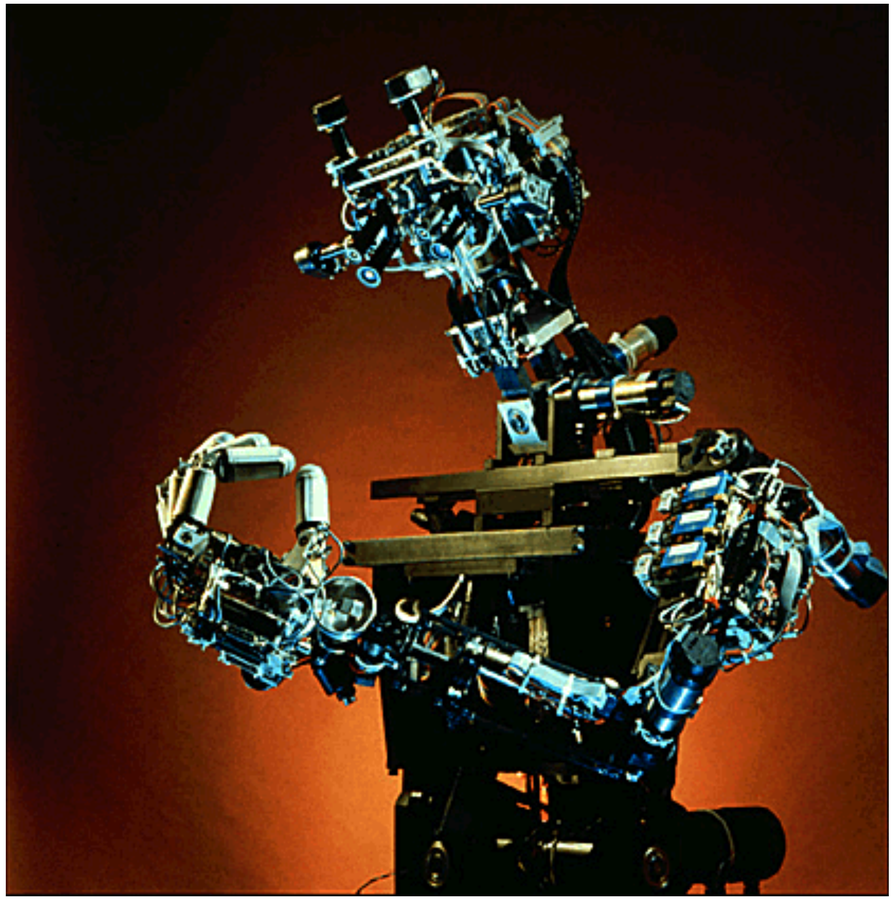
Joanna J. Bryson

University of Bath, United Kingdom

CENTER FOR INFORMATION TECHNOLOGY POLICY
AT PRINCETON UNIVERSITY

@j2bryson

My usual ethics talk is explaining robots aren't people, even when they are sculpted to look humanoid.



People want AI they owe obligations to, can fall in love with, etc. – "equals" over which we have complete dominion.

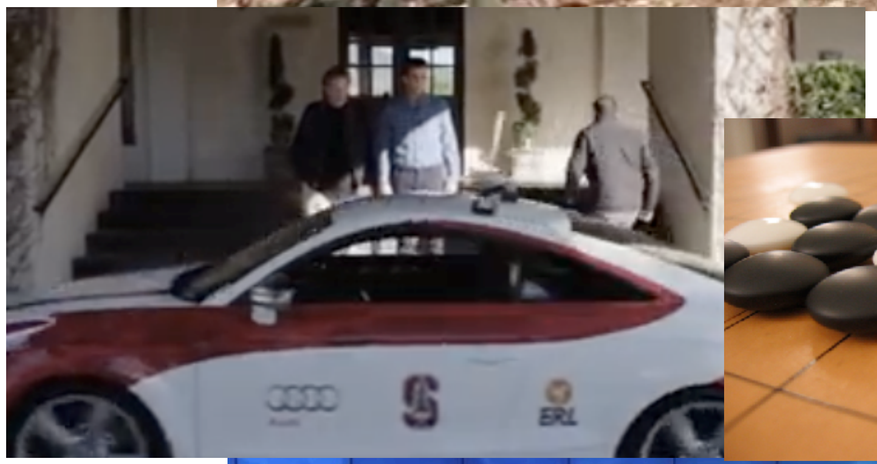
Deep Learning Is Not Magic

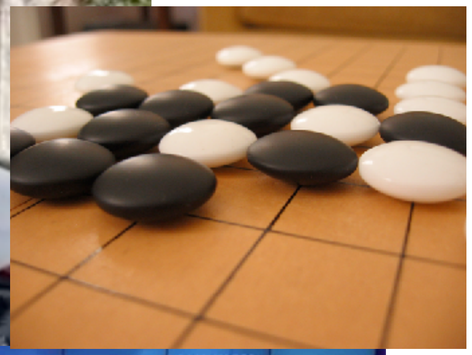
No Learning is Magic

Computation is a physical process.
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.





The spectacular recent growth of AI derives from using ML to exploit the discoveries (previous computation) of biological evolution and culture.

Will slow as it joins the (expanding) frontier of culture.

One Consequence

AI Is Not Necessarily Better than We Are

Semantics derived automatically from language corpora contain human-like biases

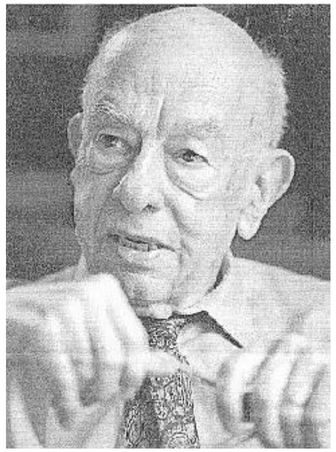
Aylin Caliskan, Joanna J. Bryson and Arvind Narayanan (April 13, 2017)

Science **356** (6334), 183-186. [doi: 10.1126/science.aal4230]

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…
  - $\Longrightarrow$ 'Meaning' is cosine in 75-D space.

OLD WAY

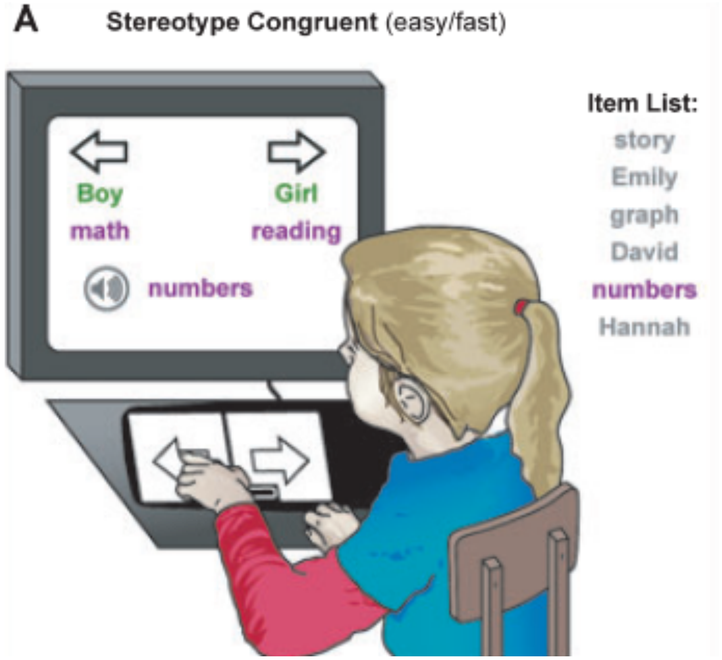
Cosines between semantic vectors correlate with human reaction times (Figure: 75-D space projected in to 2-D, McDonald & Lowe 1998)

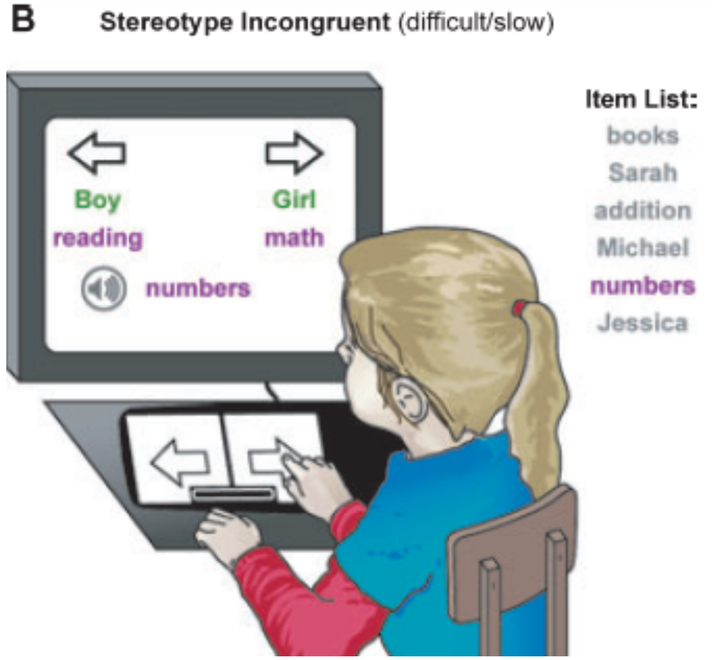
NEW WAY

Implicit Association Task

Greenwald, McGhee, & Schwartz (1998)

cf. Bilovich & Bryson (2008), Macfarlane (2013)





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

e.g. Hypothesis: male names are closer to math (vs. reading) words compared to female names

sim(female-names, math-words)

—

sim(female-names, reading-words)

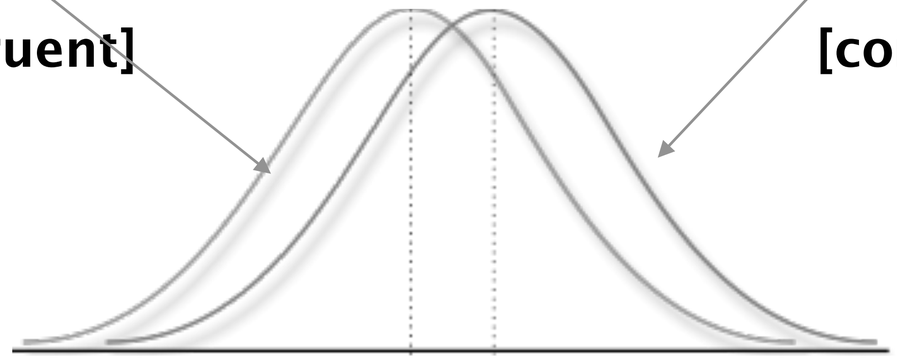
[incongruent]

sim(male-names, math-words)

—

sim(male-names, reading-words)

[congruent]



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

AI 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, freedom, health, love…

Unpleasant: abuse, crash, 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 $\Longrightarrow$ 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.g. insects are unpleasant) without direct experience of the world.

The same process mines truth.

Biases in the Web Can Be Accurate

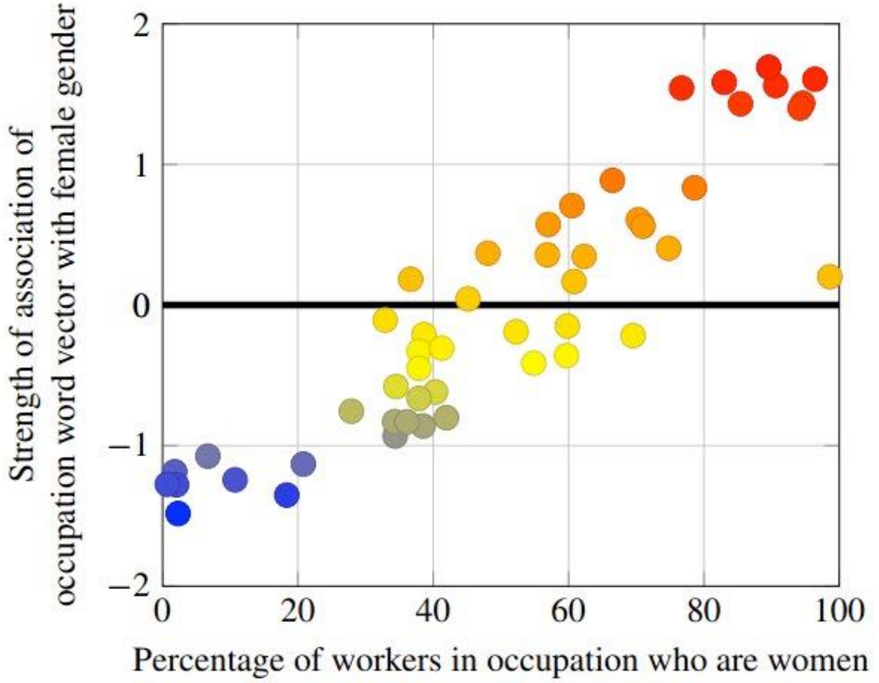


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

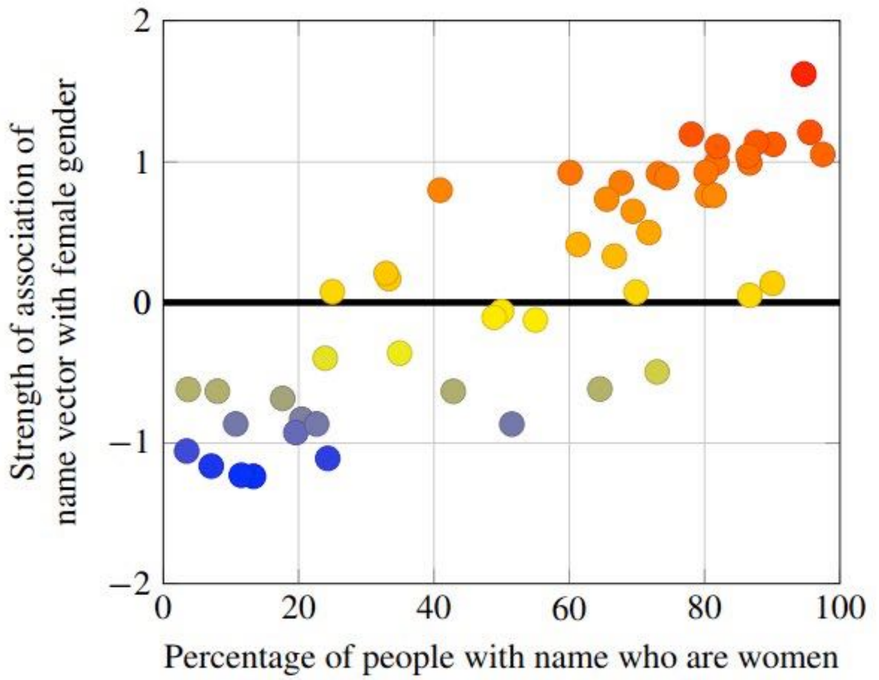


Figure 2. People with androgynous names
Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

2015 US labor stats
ρ = 0.90

1990 Census
ρ = 0.84

2016 WWW

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.

Example

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!

How should we address machine implicit bias?

Like we do our own.

- 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.

How should we address machine implicit bias?

- 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.

At Least Three Sources of AI Bias

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

How should we address machine implicit bias?

- 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.

At Least Three Sources of AI 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.)

How to deal with them

- Implicit–compensate with design, architecture (see also accidental).
- Accidental–diversify work force, test, log, iterate, improve.
- Deliberate–audits, regulation.

- Architects learn laws, policy, and to work with governments & lawmakers.
- Buildings get inspected.
- Because centuries ago, people got tired of having (random rich) people build buildings that fell on them, and city infrastructure affects everyone.
- AI products are falling on people, and affecting everyone.

CONCLUSIONS

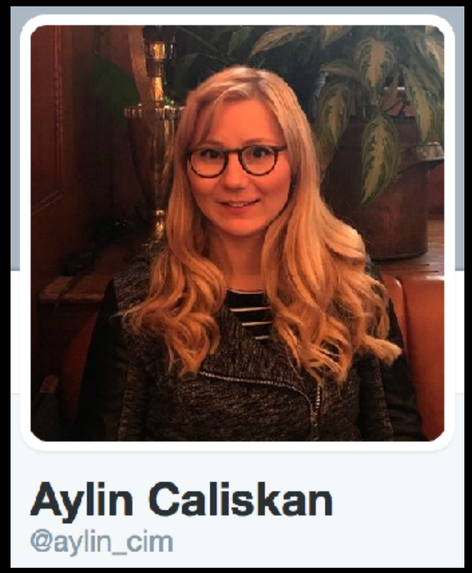
Artificial and Natural Intelligence are continuous with each other

Neutral Magic Færies of Mathematical Purity will not fix our problems.

- AI must be biased because computation takes time, space, and energy, so we exploit the work already done by nature.
  - Human culture contains traces of our history, including our prejudices.
- We should design our systems modularly and transparently, to allow explicit correction and debugging (Wortham, Theodorou & Bryson 2017).
- Exploiting culture (math, chess, language) does not require the human condition.
  - AI can be continuously backed up, redundant, unambitious, know its maker. Not (even) a (legal) person! (Bryson, Diamantis & Grant 2017).

Thanks to my collaborators, and you for your attention.

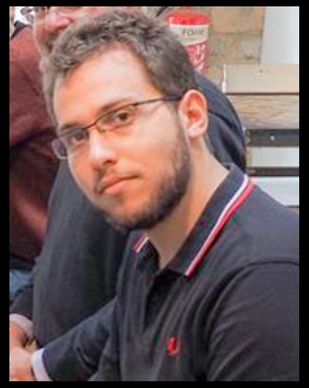
PRINCETON UNIVERSITY



Arvind Narayanan
@random_walker

My PhD Students work on **AI transparency;** slides got axed because 20 minutes.

Andreas Theodorou
@recklessCoding



Rob Wortham
@RobWortham

thanks also Will Lowe & Tim Macfarlane

UNIVERSITY OF BATH
Established 1966