Should we let someone use AI to delete human bias? Would we know what we were saying?

July 28, 2016

Update: a draft of our paper is now available too (as of 24 August), see related brief blogpost, Semantics derived automatically from language corpora necessarily contain human biases. *This work is now in published in Science as of 14 April 2017, see links below.*

As some of you will know, my colleagues and I got somewhat scooped. Aylin Caliskan, Arvind Narayanan and I are sitting on a ton of results showing that standard Natural Language tools have the same biases as humans. This is a huge deal for ethics and cognitive science, because it means that children also could learn bias just by learning language, though with children we can also give them explicit instructions to consider every human to be just as important as they are. Hopefully a draft of our paper will be available soon in arxiv.

However, even my computer scientist friends on Facebook are sharing an article Tech Review wrote about the phenomenon as it was revealed by some awesome work by Microsoft & BU. We had heard about that effort a couple months ago – those guys have been working on this for years. Of course, so have I, I've been giving talks about this model of semantics since 2001, and have published a few papers about it, notably Embodiment *vs.* Memetics (pdf, from Mind & Society, **7**(1):77-94, June 2008), but also some work on how it relates to human biases I've done with undergraduates. Aylin and Arvind have moved that work way further than I could have on my own, even with the amazing students we get at Bath, and I look forwards to sharing what we've done soon.

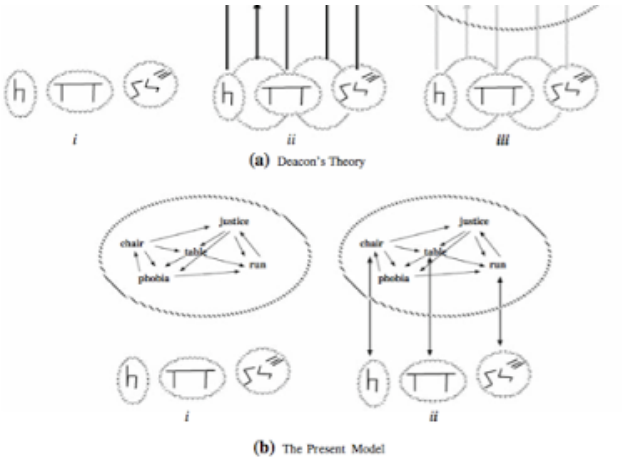


Fig. 2 In Deacon's theory, first concepts are learned [a(i)], then labels for these concepts [a(ii)], then a symbolic network somewhat like semantics$_{o/w}$ [a(iii)]. I propose instead that grounded$_{o/w}$ concepts and semantics$_{o/w}$ are learned in parallel [b(i)], with some semantic$_{o/w}$ terms becoming understood$_{o/w}$ [b(ii)]

From Embodiment vs Memetics (my 2008 paper).
We learn what words mean by how they are used, then link some to concepts we've acquired from experience.

But I'm blogging now because Bolukbasi &al. were less interested in the cognitive science of the bias in AI ethics and more interested in *fixing* it. They have used machine learning techniques to automatically suppress the links that our culture has given to historical regularities like which industries were willing to employ women.

This is a fascinating can of worms to open, and one that Aylin, Arvind & I had been discussing too. If AI generates language using meanings that do not come directly from any human culture, then on the one hand that might bring about positive change. We are language-absorbing machines; enough examples of "good" usage might change the way we talk and to some extent think. **But who do we want to pick what our language is getting changed to?** Traditionally language change has been effected mostly implicitly, by evolution-like changes driven by both fashion and necessity, but also explicitly by influential academics, textbook and newspaper editors and publishers, and even government bodies like the Académie française. Do we want new norms set now by technology companies? Would such interventions be regulated by law?

Another problem: **if AI tools don't use language like we do, it will necessarily make AI more alien and harder to understand**. Though maybe it's a good thing for a machine to be conspicuously a machine, this is a form of transparency. But if the technology is being used to communicate in critical situations, it might be better to make it as comprehensible as possible.

So far, I think my coauthors and I have been more focussed on using technology to encourage **human**

That's a big difference. I look forwards to the debate! But right now I'm getting back to the papers I'm writing.

Related publications:

- Just last week Aylin won a prize for best talk on this work in the "hot takes" section of PET (a privacy meeting). Here's her abstract.
- Embodiment *vs.* Memetics (pdf, from Mind & Society, **7**(1):77-94, June 2008) was actually first presented as aposter at a workshop at CogSci 2001, and was also was a talk at Evolution of Language 2003.
    - Since then I've updated my theory of language evolution and human uniqueness (that's a blogpost summarsing), and
    - I also gave a plenary at AISB 2015 about this model of semantics impacts AI ethics Embodiment vs Memetics: From Semantics to Moral Patiency through the Simulation of Behaviour (that's the slides from the meeting in PDF.)
- My first attempt at applying this to human bias was only semi-successful but still cool: Detecting the Evolution of Semantics and Individual Beliefs Through Statistical Analysis of Language Use, Bilovich & Bryson, Proceedings of the Fall AAAI Symposium on Naturally-Inspired Artificial Intelligence, Washington DC, November 2008.
- I since had a more successful attempt but with an undergraduate who didn't have time to publish... the 2013 tech report is here: http://opus.bath.ac.uk/37916/

Update April 2017 The MS/BU paper and ours hit arxiv about the same time last year. The MS/BU one ultimately got into NIPS. *Our* paper is now in Science. I just found out about another related paper that came out in Cognition a few weeks ago.

Other AiNI blogposts on this work:

- Should we let someone use AI to delete human bias? Would we know what we were saying? 28 July 2016
- Semantics derived automatically from language corpora necessarily contain human biases 24 August 2016
- FAQ for our Semantics paper 13 April 2017
- We Didn't Prove Prejudice Is True (A Role for Consciousness) 13 April 2017

3 comments

Add a comment

Top comments

Joanna Bryson 8 months ago - Shared publicly

We were discussing at lunch today who to send our paper to **before** we upload it to arxiv, and we realised we could use our software to predict how receptive they might be. I **think** we were joking.

In other news, we emailed this blogpost when I wrote it to the Microsoft / BU crowd as well as linking to them here, and two of them replied to my email, but they posted their own blog a week ago without linking to this.

1 · Reply

Joanna Bryson via Google+ 9 months ago - Shared publicly

AI derived from human culture will be as biased as our history is. What should we do?

1 · Reply

Gordon Mohr 9 months ago (edited) - Shared publicly

I'm somehow reminded of an exchange from 'Eternal Sunshine of the Spotless Mind':

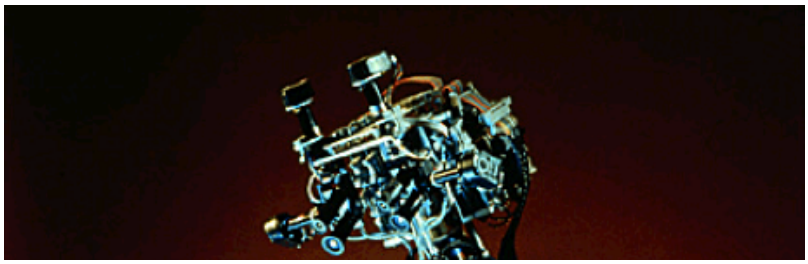
JOEL: Is there any risk of brain damage?
DOCTOR: Well, technically speaking, the operation **is** brain damage… but it's on a par with a night of heavy drinking. Nothing you'll miss.

+1 1 · Reply

0

Why (or rather, when) suffering in AI is incoherent.

I've been arguing for some months now *in public talks that AI cannot be a legal person because suffering in well-designed AI is incoherent. This is not actually my own argument, but rather is due to S. M. Solaiman from their brilliant recent article* Legal personality of robots, cc ...



Robots are owned. Owners are taxed. Internet services cost Information.

As I often recount, I got involved in AI ethics *because I was dumbfounded that people attributed moral patiency to (thought I shouldn't unplug)* Cog*, the humanoid robot, when in fact it wasn't plugged in, and didn't work (this was 1993-1994). The processors of* ...

Semantics derived automatically from language corpora necessarily contain human biases

Here is a draft of the paper I promised last month:

Aylin Caliskan-Islam*, Joanna J. Bryson, &* Arvind Narayanan*,* Semantics derived automatically from language corpor ...