FAQ for our Semantics paper

April 13, 2017

These are FAQs about our paper, Semantics derived automatically from language corpora contain human biases. My main blogpost about that is We Didn't Prove Prejudice Is True (A Role for Consciousness). There are some other blogposts at the bottom of the page here.

Q: So words that mean similar things co-occur, right?

Semantically similar words *may* sometimes cooccur, but the main point of word embeddings is that two semantically similar words tend to occur in similar contexts, which means the other words around them are similar. Like "I need to go home and feed my" occurs for both *cat* and *dog*, but not *justice*.

Q: Doesn't GloVe already have word embeddings? What does WEAT add? What's the point?

The point of our paper is really that we have thought of asking the question of whether implicit associations (including the ones associated with prejudices) can be communicated through word use – through semantics. Well, first that we asked the question, second that the answer was "yes", and third the answer was so firm.

WEAT is basically a methodology for checking how sure we are that two sets of words are more associated with each other than they are with two their opposite partners. So this is a methodology we run on top of the word embeddings that have been built using GloVe (in the main article), word2vec (in the supplement), or probably any other good word embedding method. We aren't creating the word embeddings; we are using preexisting word embeddings and seeing whether they contain information concordant with the stereotypes documented with the Implicit Association Test (IAT).

So in a way, your question is like asking whether the IAT adds anything on top of a set of people. WEAT, like IAT, is a test. Running GloVe over the CommonCrawl corpus gives us the "subject" word embeddings that we run the WEAT over. What's interesting is that the WEAT test of word embeddings gives an analogous to result of the IAT test of American people.

Computers don't do anything. They are artefacts. People and organisations do things with computers.

But if we change your question to be "why would AI contain human prejudices?", the reason is because computation is not math – it's not a pure abstraction with certain truth. **Computation is a physical process that happens in the real world. It takes time, energy, and space.** The reason humans are so much smarter than other apes is because we have gotten very good at passing the outcomes of our computation to each other. In the last ten years we have gotten very good at passing the outcomes of *our* computation *to each other* through AI, and even on extending our intelligence using AI to do automated search. What our paper really shows is that the biases of the IAT are a part of our implicit culture. They are part of the information we pass each other when we talk to each other. Aylin, Arvind and I are just using AI to measure this.

Q: Isn't it horrible that we might get biases from talking to AI?

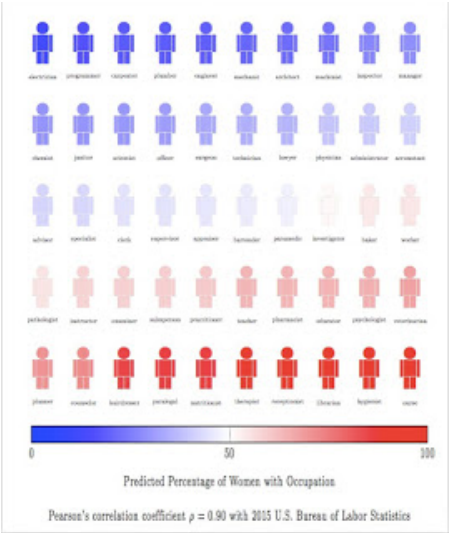
It's exactly as horrible as the biases we get by talking to each other. You can't talk without using language. So long as we keep updating the word embeddings, then as we figure out ways to improve our society, we will also improve our AI at the same time.

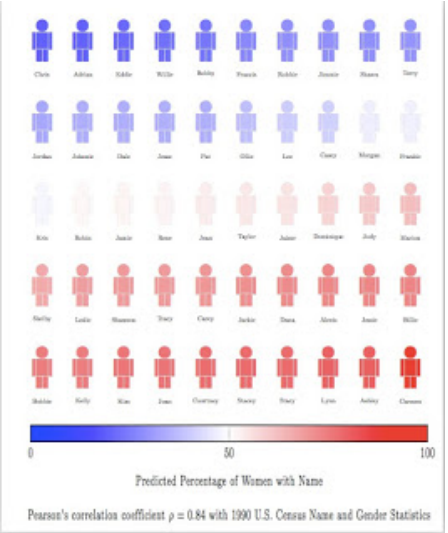
Q: If it's in a machine, can't we just fix it?

1. This isn't just about prejudice. Bias is how we know anything at all. We can't change what language is in the real world, and we wouldn't really want to if that meant we couldn't communicate anymore.
2. If we could change language, who would get to decide how it should be? (I wrote that blogpost back right before we first put this work in arxiv.)
3. Well, yes. We could fix it exactly how humans fix it. We could build AI language systems with a more cognitive architecture, using both implicit and explicit memory. We could use word embeddings as implicit memory to understand and represent language that is heard, and even for creating draft possible sentences. But then we could use explicit rules that would be open to human examination to filter what actually gets said. We could recognise terms that might be problematic, like "the typical programmer likes to tie his tie in a knot" and apply a transform like "the typical programmer likes to tie their tie in a knot."

Q: I just want to know what the names and jobs were in your WEFAT results.

I'm so glad you asked! Fortunately Aylin Caliskan made these lovely figures:





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

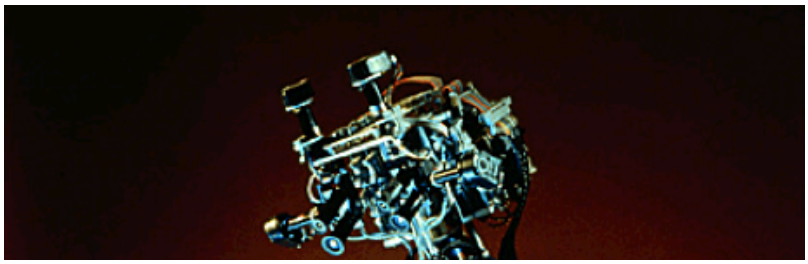
AI ETHICS SCIENCE

Add a comment

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 ...