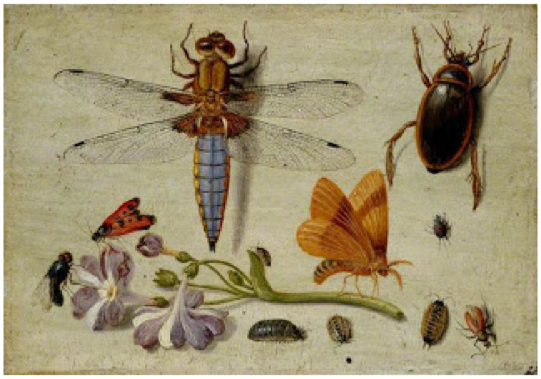
We Didn't Prove Prejudice Is True (A Role for Consciousness)

April 13, 2017

The good news: We know where word meanings come from



Paintings due to Jan Van Kessell

We have a paper in *Science*, Semantics derived automatically from language corpora contain human biases (a green open access version is hosted at Bath). What this paper shows is that you can find the implicit biases humans have just by learning semantics from our language. We showed this by using machine learning of semantics from the language on the Web, and comparing that to implicit biases psychologists have documented using the Implicit Association Test (IAT). The IAT uses reaction times to show that people find it easier to associate some things than others. For example, it's easier to associate flowers with pleasant terms and bugs with unpleasant terms than the other way around. Notice that the actual statistics underlying the IAT is always about these slightly complicated, dual *relative* measures. It's easier to:

> *group {flowers and pleasant terms} together, AND {unpleasant terms and insect names} (both those groupings at once), than to do the opposite two groupings: {flowers and unpleasant terms}, AND {pleasant terms and insect names.}*

People tend to simplify this by saying things like "people find **flowers** more pleasant than insects", but that's not entirely correct. You can't really take out just part of an IAT like that.

What we did was develop an equivalent test to the IAT for the meanings of words, where meaning is defined by the way words are used. What I mean by "the way words are used" is the contexts you find the words in. For example, the word *flowers* I highlighted just above is occurring in a context of "saying things like 'people find" and "more pleasant than insects', but". Those are just the five words on either side of the target word, *flowers*. When you collect together a whole bunch of such contexts and compress the results (compression gives you generalisation) it's called a word embedding.

how it is used – that in fact usage could be a valid definition for what *meaning* means. Some people have argued this for a while, and in fact it's basic to how search engines work: latent semantic analysis based on this kind of meaning is why you can find a whole website from just a few words. It may seem like a weird reductionist or even cyclic definition of *meaning* but it's not really cyclic, and another word for reductionist is "elegant". I've been excited by this idea for decades, in fact I married the person who first explained it to me.

What our paper shows that is new is that this implies that you can know visceral facts about the real world having no other experience of it than being exposed to the regularities of human language. **Word embeddings seem to know that insects are icky and flowers are beautiful.** We haven't tested exactly that, but that's the implication of sharing the same kind of biases humans do, which is what we showed. This is amazing, I think it's probably worth a paper in *Science* itself. It goes completely against the theories of embodiment that my initial PhD on Cog the robot were based on. **Our result shows humans can give each other understanding not only by logic and argument, but just by how we talk.** Even if you don't understand sentences, you can pick up regularities about the nature of terms that describe the natural world just by keeping track of how and where words are used. That's incredibly important, because **we need some way to bootstrap (kick start) the system by which we do argue.** We need to know words' meanings before we can start using them to argue and explain. Now my colleagues and I have shown a model for how that might be accomplished**.

The bad news: AI (and children) can inherit our prejudices just from word usage

Of course the headline-grabbing thing about our work was not the flowers and insects. What the IAT is better known for is showing the extent to which we have strong implicit associations for a wide range of stereotypes, like that women are more domestic an men more career oriented, or women are more associated with the humanities and men to math or science. And worst of all, that African American first names are more easily associated with unpleasant terms AND European American names with pleasant terms than the other way around. We showed that all these associations are present in automatically generated word embeddings as well.

Another thing we show is that **the same representations that produce the stereotyped associations for women and men are highly correlated with real facts about men and women**, like what proportion of credible hypothesis that **stereotypes are just the regularities that exist in the real world that our society has decided we want to change**.

Some implicit biases are objectively false

The most terrible thing anyone has said to me about this work is "You realise you just proved prejudices are true." No, we just showed that some biases reflect some aspects of historic reality. One of the stereotypes I've seen demonstrated by one of the leading IAT researchers is that we associate {{good with our right hands} and {bad with our left hands}} way more easily than we associate {{good with our left hands} and {bad with our right hands.}} Next to no one has believed that the left side of our body was bad in Europe for centuries (the Romans believed it!) You can't say that it's true. But it's a very strong implicit bias, because it's been a big part of our historic culture.

In the very first paper with the IAT, Greenwald and his colleagues also tested the association between pleasantness and unpleasantness and Korean and Japanese names. They did that in populations of Korean Americans and populations of Japanese Americans. Both groups found the other group's names more easily associated with unpleasant and their own more easily associated with pleasant terms. What this probably means is that it's pleasant to be in-group, with familiar things.

We couldn't use WEAT to try to find a similar result, because we couldn't tell what part of the Web was written by Korean Americans or Japanese Americans. But IAT researchers have shown that unfortunately, African Americans share the same implicit bias as European Americans. But this doesn't mean that African Americans are unpleasant any more than the IAT means left is really bad. What it probably means is something like that the names of people who dominate a particular culture are somehow associated more with some pleasant things. We have a few ideas of how to get at what this means better, but we haven't had time to try them yet. I hope we'll have another paper about this within a year, but there's no way to promise that.

A use for consciousness

In the mean time, one of the amazing things all this has brought home to me is that there's a very good reason to have an architecture with both implicit and explicit memory. The decision by our culture to let rely only on what has been historically true. But we do, as I say, need to bootstrap off of what has been historically true -- by and large, it's useful to know about observed facts in the world.

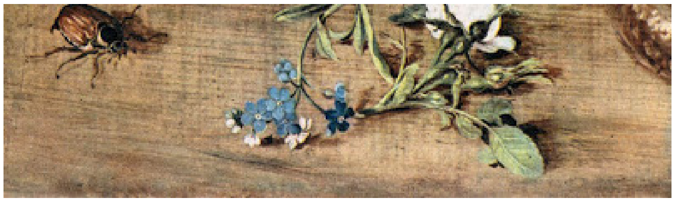
So if we are going to have AI systems that learn about the world from culture and then also act upon the world, then we may very well want a similar system to what humans have. We may want an implicit system for learning enough scaffolding to understand the world, and then an explicit way of instructing the system to conform to what society currently accepts. This may sound scifi, but think about the text prediction on your own smart phone (if you have one.) It guesses the next word you might type with something derived from culture -- an n-gram model that tells it what words you are likely to say next, particularly given what letters you've typed so far. But there are some words it will never help you finish. That's not because no one has ever said them before, it's because guessing them wrong would be socially unacceptable.

So we already have AI systems that are cognitive in this simple way. They work off of general cultural precedent, except when that violates a politeness that was worth programming into them separately.

And more excitingly, we can see a reason for why it might be useful to split human intelligence into two systems; implicit and explicit. The one we have conscious access to is the one we use for negotiating new societies and making progress. The one we don't have conscious access too learns all kinds of other regularities to let the conscious one get stuff done. Just a hypothesis, but a cool one.

Footnote

** Normally I say "models aren't data about the world, they're just data about theories." Though in this case I guess you could say we have gotten data about how language could be used. But *could* is the operative word. Science doesn't prove things; proofs are for abstractions like logic and math. Science builds better and better models of the world, but models by their definition are (also) abstractions and therefore in some sense "wrong" -- they aren't the world. So anyway, we haven't shown that children do learn prejudice this way, we've shown that they could. But the fact that they could does increase the probability that they do.



You can see some Van Kessel paintings at the Holburne Museum
if you come to Bath soon, e.g. for AISB 2017 18-21 April.

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

Update 17 April: this stuff is clearly zeitgeist; besides the two arxiv articles last summer, other people with similar things are getting in touch

- My own earlier papers on this project (going back to 2001!) are listed in my first blogpost last summer about this new collaborative effort with Aylin & Arvind: Should we let someone use AI to delete human bias? Would we know what we were saying? which also discusses the Bolukbasi &al. paper that hit arxiv the same time ours did.
- A paper on the IAT & word embeddings got accepted by *Cognition* 14 days after ours was accepted by *Science*; it won't be published in print until July, but has actually been on line a few days longer than ours: The semantic representation of prejudice and stereotypes.
- There was also a related CogSci paper in 2012, based on n-grams not word embeddings. Modelling the IAT : Implicit Association Test reflects shallow linguistic environment and not deep personal attitudes

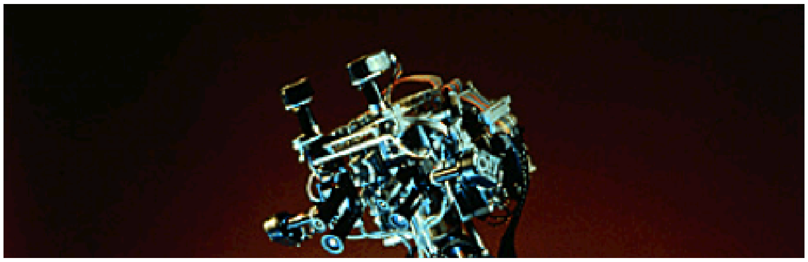
AI ETHICS SCIENCE

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, co …



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 …