DailyNous | July 30, 2020 | "Philosophers on GPT-3" series, edited by Annette Zimmermann
Contributions by Amanda Askell, David Chalmers, Justin Khoo, Carlos Montemayor, C. Thi Nguyen, Regina Rini, Henry Shevlin, Shannon Vallor, and Annette Zimmermann

If You Can Do Things with Words, You Can Do Things with Algorithms

Annette Zimmermann

Ask GPT-3 to write a story about Twitter in the voice of Jerome K. Jerome, prompting it with just one word ("It") and a title ("The importance of being on Twitter"), and it produces the following text: "It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage." Sounds plausible enough—delightfully obnoxious, even. Large parts of the AI community have been nothing short of ecstatic about GPT-3's seemingly unparalleled powers: "Playing with GPT-3 feels like seeing the future," one technologist reports, somewhat breathlessly: "I've gotten it to write songs, stories, press releases, guitar tabs, interviews, essays, technical manuals. It's shockingly good."

Shockingly good, certainly—but on the other hand, GPT-3 is *predictably bad* in at least one sense: like other forms of AI and machine learning, it reflects patterns of historical bias and inequity. GPT-3 has been trained on us—on *a lot* of things that we have said and written—and ends up reproducing just that, racial and gender bias included. OpenAI acknowledges this in their own paper on GPT-3,[1] where they contrast the biased words GPT-3 used most frequently to describe men and women, following prompts like "He was very…" and "She would be described as…". The results aren't great. For men? Lazy. Large. Fantastic. Eccentric. Stable. Protect. Survive. For women? Bubbly, naughty, easy-going, petite, pregnant, gorgeous.

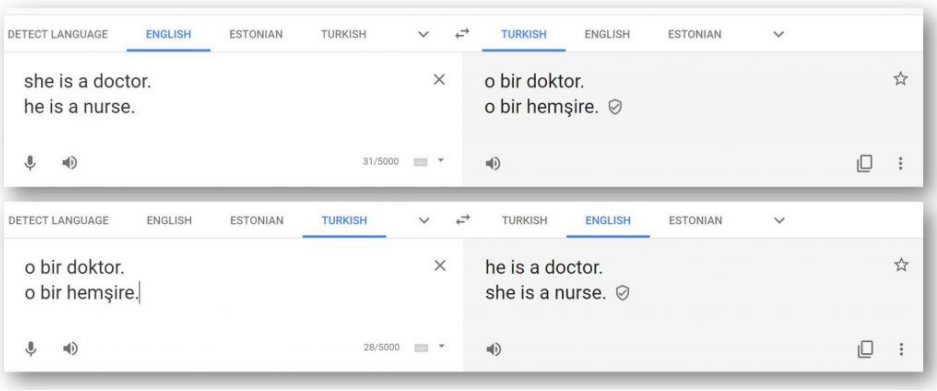
These findings suggest a complex moral, social, and political problem space, rather than a purely technological one. Not all uses of AI, of course, are inherently objectionable, or automatically unjust—the point is simply that much like we can do things with words, we can *do* things with algorithms and machine learning models. This is not purely a tangibly material *distributive justice* concern: especially in the context of language models like GPT-3, paying attention to other facets of injustice—*relational, communicative, representational, ontological*—is essential.

Background conditions of structural injustice—as I have argued elsewhere—will neither be fixed by purely technological solutions, not will it be possible to analyze them fully by drawing exclusively on conceptual resources in computer science, applied mathematics and statistics. A recent paper by machine learning researchers argues that "work analyzing "bias" in NLP systems [has not been sufficiently grounded] in the relevant literature outside of NLP that explores the relationships between language and social hierarchies," including philosophy, cognitive linguistics, sociolinguistics, and linguistic anthropology. Interestingly, the view that AI development might benefit from insights from linguistics and philosophy is actually less novel than one might expect. In September 1988, researchers at MIT published a student guide titled "How to Do Research at the MIT AI Lab", arguing that "[l]inguistics is vital if you are going to do natural language work. […] Check out George Lakoff's recent book *Women, Fire, and Dangerous Things*." (Flatteringly, the document also states: "[p]hilosophy is the *hidden framework* in which *all AI* is done. Most work in AI takes implicit philosophical positions without knowing it").

Following the 1988 guide's suggestion above, consider for a moment Lakoff's well-known work on the different cognitive models we may have for the seemingly straightforward concept of 'mother', for example: 'biological mother', 'surrogate mother', 'unwed mother', 'stepmother', 'working mother' all denote motherhood, but neither one of them picks out a socially and culturally uncontested set of necessary and sufficient conditions of motherhood.[3] Our linguistic practices reveal complex and potentially conflicting models of who is or counts as a mother. As Sally Haslanger has argued, the act of defining 'mother' and other contested categories is subject to non-trivial disagreement, and necessarily involves implicit, internalized assumptions as well as explicit, deliberate political judgments.[4]

Very similar issues arise in the context of all contemporary forms of AI and machine learning, including but going beyond NLP tools like GPT-3: in order to build an algorithmic criminal recidivism risk scoring system, for example, I need to have

a conception in mind of what the label 'high risk' *means*, and how to measure it. Social practices affect the ways in which concepts like 'high risk' might be defined, and as a result, which groups are at risk of being unjustly labeled as 'high risk'. Another well-known example, closer to the context of NLP tools like GPT-3, shows that even words like gender-neutral pronouns (such as the Turkish third-person singular pronoun "o") can reflect historical patterns of gender bias: until fairly recently, translating "she is a doctor/he is a nurse" to the Turkish "o bir doktor/o bir hemşire" and then back to English used to deliver: "*he* is a doctor/*she* is nurse" on GoogleTranslate.[5]



[source]

The bottom line is: social meaning and linguistic context matter a great deal for AI design—we cannot simply assume that design choices underpinning technology are normatively neutral. It is unavoidable that technological models interact dynamically with the social world, and vice versa, which is why even a perfect technological model would produce unjust results if deployed in an unjust world.

This problem, of course, is not unique to GPT-3. However, a powerful language model might *supercharge* inequality expressed via linguistic categories, given the scale at which it operates.

If what we care about (amongst other things) is *justice* when we think about GPT-3 and other AI-driven technology, we must take a closer look at the linguistic categories underpinning AI design. If we can politically critique and contest social practices, we can critique and contest language use. Here, our aim should be to engineer conceptual categories that mitigate conditions of injustice rather than entrenching them further. We need to deliberate and argue about which social practices and structures—including linguistic ones—are morally and politically valuable *before* we automate and thereby accelerate them.

But in order to do this well, we can't just ask how we can *optimize* tools like GPT-3 in order to get it closer to humans. While benchmarking on humans is plausible in a 'Turing test' context in which we try to assess the possibility of machine consciousness and understanding, why benchmark on humans when it comes to creating a more just world? Our track record in that domain has been—at least in part—underwhelming. When it comes to assessing the extent to which language models like GPT-3 moves us closer to, or further away, from justice (and other important ethical and political goals), we should not necessarily take ourselves, and our social status quo, as an implicitly desirable baseline.

A better approach is to ask: what is the *purpose* of using a given AI tool to solve a given set of tasks? How does using AI in a given domain shift, or reify, power in society? Would redefining the problem space itself, rather than optimizing for decision quality, get us closer to justice?

Notes

(1) Brown, Tom B. et al. "Language Models are Few-Shot Learners," arXiv:2005.14165v4.
(2) Blodgett, Su Lin; Barocas, Solon; Daumé, Hal; Wallach, Hanna. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP," arXiv:2005.14050v2.
(3) Lakoff, George. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press (1987).
(4) Haslanger, Sally. "Social Meaning and Philosophical Method." *American Philosophical Association 110th Eastern Division Annual Meeting* (2013).
(5) Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science* 356, no. 6334 (2017), 183-186.