Abstract for

Algorithmic Bias: on the implicit biases of social technology

by Gabbrielle M Johnson

gmjohnson@humnet.ucla.edu

UCLA, Philosophy

December 18th, 2017

word count: 912

Algorithmic Bias: on the implicit biases of social technology

On March 23, 2016, Microsoft released Tay, an artificial intelligence (AI) Twitter chatbot intended to interact with other Twitter users and mimic the language patterns of a 19-year-old American girl. Tay operated by learning from the language patterns of human Twitter users with whom it interacted. Within 16 hours of its launch, Tay had authored a number of tweets endorsing Nazi ideology and harassing other Twitter users. When asked about how Tay developed such a noxious personality, Microsoft responded "As [Tay] learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it."[1] In other words, Tay's personality was inherited from the individuals with whom it was engaging.

Tay's story highlights an obstacle facing developers of machine learning programs: *implicit algorithmic bias*. An AI like Tay, which uses machine learning to capitalize on (or "learn" from) statistical regularities in human-generated data-sets, tends to pick up on social patterns that manifest in human behavior and that are reflected in the data on which it is trained. In cases like Tay's, we have reason to suspect that programmers are not explicitly writing biases toward marginalized demographics into their software's code.[2] Instead, the biases appear to *implicitly emerge* from the algorithms' operating on the data, mimicking the biases reflected in the data themselves. This possibility raises troubling questions about the relationship between algorithmic and cognitive biases, as well as their relationship to overarching structural and systemic biases.

This paper explores the relationship between algorithmic and cognitive biases and argues that they are of the same basic kind. I begin by presenting cases where algorithmic biases mimic patterns of human implicit bias. As a clear example, consider a study by Caliskan et al. (2017) on word-embedding machine learning.[3] This study found that parsing software trained on a dataset of 840 billion words collected by crawling the internet resulted in the program producing "human-like semantic biases" that replicated well-known trends in human implicit bias tests. These biases included the tendency to more often pair stereotypical female names with family words than career terms, stereotypical African-American names with unpleasant words rather than pleasant words, and stereotypical male names with science and math terms rather than art terms. On this result, co-author Arvind Narayanan writes, "natural language necessarily contains human biases, and the paradigm of training machine learning on language corpora means that AI will inevitably imbibe these biases as well."[4]

Because algorithmic biases don't utilize explicitly represented bias rules, cases such as Tay's are not naturally accommodated by extant models of implicit cognitive bias, which either fail to fully consider or mischaracterize biases whose components are not explicitly represented. In contrast, the functional model of implicit bias that I develop in previous work accommodates both representational and non-representational biases. In this paper, I argue that my functional model has two advantages over these other accounts.

---

[1] **Price, R. (2016). "Microsoft is deleting its AI chatbot's incredibly racist tweets."** ***Business Insider*****.**

[2] See, for example, O'Neil's discussion of an event where Google's automatic photo-tagging service labeled a group of African American's as gorillas. **O'Neil, C., p. 154, (2016).** ***Weapons of Math Destruction: How big data increases inequality and threatens democracy*****. Crown Publishing Group.**

[3] **Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). "Semantics derived automatically from language corpora contain human-like biases."** ***Science*****, 356(6334):183–186.**

[4] **Narayanan, A. (2016). "Language necessarily contains human biases, and so will machines trained on language corpora."**

First, models of implicit bias take human cognitive bias as the paradigmatic case; but, it may turn out that some human cognitive bias is, like Tay's algorithmic bias, non-representational.[5] To motivate this possibility, I demonstrate the point using a simple toy model of the *k*-nearest neighbors (KNN) learning algorithm which serves as a proof of concept: some algorithms operate *as if* they contain explicit reference to stereotypes even when in fact they don't. Next, I argue that the same may apply to human implicit bias, i.e., some cognitive biases influence an individual's beliefs about and actions toward other people, but are, nevertheless, nowhere represented in that individual's cognitive repertoire. I call these *truly implicit biases*. I then demonstrate how the aforementioned functional model of implicit bias can be straightforwardly extended to handle both cases of non-represented biases.

Second, my functional account defines implicit bias in terms of propositional inputs and outputs (represented in the algorithmic case as combinations of objects, feature-values, and class labels), and so has the advantage of allowing for robust predictive and explanatory exchange between the algorithmic and cognitive domains, independent of whether the biases of these domains are representational.[6] I conclude by using the comparison of these two cases to demonstrate one plausible explanation for why human implicit biases resist revision. In cases of algorithmic biases, programmers have long struggled with the difficulty of eliminating biases that are based on so-called 'proxy attributes'. These are seemingly innocuous attribute labels that correlate with socially sensitive attributes, serving as proxies for the socially-sensitive attributes themselves. For example, in the historic cases of discriminatory redlining, zip codes were used as proxies for race. Crucially, the effects of these proxy attributes tend to resist any overt filtering techniques. Eliminating any explicit references to race in a program's code will not ameliorate the harms it causes since the program can simply substitute a proxy attribute in place of race, resulting in similar discriminatory effects. Likewise, one might think that human implicit biases similarly resist revision since most attempts to revise them focus on overt, socially-sensitive attributes rather than potential proxy attributes. Better understanding of these attempts to identify the operation of proxy attributes in the algorithmic case will plausibly lend to a better understanding of mitigation techniques for human implicit biases. An account of implicit bias on which both algorithmic bias and human cognitive bias are of the same basic kind anticipates and facilitates the fruitfulness of these comparisons.

---

[5] Prominent models that fall into this category include any theory that posits representations at the core of a bias's operation. These include various associative accounts including the Associative-Propositional Model (APE) from **Gawronski, B. and Bodenhausen, G. V. (2014). "Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model: APE Model."** ***Social and Personality Psychology Compass*****, 8(8):448–462**; the *alief* account from **Gendler, T. S. (2008). "Alief and belief."** ***The Journal of Philosophy*****, 105(10):634–663**; the Minimal Model from **Holroyd, J. (2016). "VIII: What Do We Want from a Model of Implicit Cognition?"** ***Proceedings of the Aristotelian Society*****, 116(2):153–179**; and the intrinsic affect-laden stereotype model from **Madva, A. and Brownstein, M. (2016). "Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind: Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind."** ***Nous***; as well as various propositional accounts including **Mandelbaum, E. (2015). "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias."** ***Nous*****, DOI:10.1111/nous.12089, p. 1–30**; **De Houwer, J. (2014). "A Propositional Model of Implicit Evaluation: Implicit evaluation."** ***Social and Personality Psychology Compass*****, 8(7):342–353**; and **Levy, N. (2015). "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements."** ***Nous*****, 49(4):800– 823**.

[6] This account borrows important insights from dispositionalist approaches to belief (e.g., those presented by **Ryle, G. (1949).** ***The Concept of Mind*****. Barnes and Noble**; **Dennett, D. C. (1981). "A Cure for the Common Code." In** ***Brainstorms: philosophical essays on mind and psychology*****, pages 90–108. MIT Press, Cambridge, Mass.**; and **Schwitzgebel, E. (2002). "A phenomenal, dispositional account of belief."** ***Nous*****, 36(2):249–275**) and, regarding models of implicit bias, most resembles the trait approach presented by **Machery, E. (2016). "De-Freuding implicit attitudes." In** ***Implicit Bias & Philosophy: Metaphysics and Epistemology,*** **volume 1, pages 104–129. Oxford University Press**. However, this view also differs from each of these approaches since, as I'll explain, it is committed to very restricted forms of disposition, namely those that are underlain by combinations of states and processes that systematically relate propositional state inputs to propositional state outputs.