Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan[1], **Joanna J. Bryson**[1,2], **and Arvind Narayanan**[1]

[1]Princeton University
[2]University of Bath
*Address correspondence to aylinc@princeton.edu, bryson@conjugateprior.org, arvindn@cs.princeton.edu.

ABSTRACT

Artificial intelligence and machine learning are in a period of astounding growth. However, there are concerns that these technologies may be used, either with or without intention, to perpetuate the prejudice and unfairness that unfortunately characterizes many human institutions. Here we show for the first time that human-like semantic biases result from the application of standard machine learning to ordinary language—the same sort of language humans are exposed to every day. We replicate a spectrum of standard human biases as exposed by the Implicit Association Test and other well-known psychological studies. We replicate these using a widely used, purely statistical machine-learning model—namely, the GloVe word embedding—trained on a corpus of text from the Web. Our results indicate that language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the *status quo* for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics. In addition to our empirical findings concerning language, we also contribute new methods for evaluating bias in text, the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). Our results have implications not only for AI and machine learning, but also for the fields of psychology, sociology, and human ethics, since they raise the possibility that mere exposure to everyday language can account for the biases we replicate here.

Introduction

Those astonished by the human-like capacities visible in the recent advances in artificial intelligence (AI) may be comforted to know the source of this progress. Machine learning, exploiting the universality of computation (Turing, 1950), is able to capture the knowledge and computation discovered and transmitted by humans and human culture. However, while leading to spectacular advances, this strategy undermines the assumption of machine neutrality. The default assumption for many was that computation, deriving from mathematics, would be pure and neutral, providing for AI a fairness beyond what is present in human society. Instead, concerns about machine prejudice are now coming to the fore—concerns that our historic biases and prejudices are being reified in machines. Documented cases of automated prejudice range from online advertising (Sweeney, 2013) to criminal sentencing (Angwin et al., 2016).

Most experts and commentators recommend that AI should always be applied transparently, and certainly without prejudice. Both the code of the algorithm and the process for applying it must be open to the public. Transparency should allow courts, companies, citizen watchdogs, and others to understand, monitor, and suggest improvements to algorithms (Oswald and Grace, 2016). Another recommendation has been diversity among AI developers, to address insensitive or under-informed training of machine learning algorithms (Sweeney, 2013; Noble, 2013; Barr, 2015; Crawford, 2016). A third has been collaboration between engineers and domain experts who are knowledgeable about historical inequalities (Sweeney, 2013).

Here we show that while all of these strategies might be helpful and even necessary, they could not be sufficient. We document machine prejudice that derives so fundamentally from human culture that it is not possible to eliminate it through

strategies such as the above. We demonstrate here for the first time what some have long suspected (Quine, 1960)—that *semantics*, the meaning of words, necessarily reflects regularities latent in our culture, some of which we now know to be prejudiced. We demonstrate this by showing that standard, widely used Natural Language Processing tools share the same biases humans demonstrate in psychological studies. These tools have their language model built through neutral automated parsing of large corpora derived from the ordinary Web; that is, they are exposed to language much like any human would be. Bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.

Human learning is also a form of computation. Therefore our finding that data derived from human culture will deliver biases and prejudice have implications for the human sciences as well. They present a new "null hypothesis" for explaining the transmission of prejudice between humans. Our findings also have implications for addressing prejudice, whether in humans or machines. The fact that it is rooted in language makes prejudice difficult to address, but by no means impossible. We argue that prejudice must be addressed as a component of any intelligent system learning from our culture. It cannot be entirely eliminated from the system, but rather must be compensated for.

In this article, we begin by explaining meaning and the methods by which we determine human understanding, and interpret it in machines. Then we present our results. We replicate previously-documented biases and prejudices in attitudes towards ordinary objects, animals, and humans. We show that prejudices that reduce the number of interview invitations sent to people because of the racial association of their name, and that associate women with arts rather than science or mathematics, can be retrieved from standard language tools used in ordinary AI products. We also show that veridical information about the proportions of women in particular job categories, or what proportion of men versus women have a particular name, can be recovered using the same methods. We then present a detailed account of our methods, and further discussion of the implications of our work.

Meaning and Bias in Humans and Machines

In AI and machine learning, *bias* refers to prior information, a necessary prerequisite for intelligence (Bishop, 2006). Yet bias can be problematic where prior information is derived from precedents known to be harmful. For the purpose of this paper, we will call harmful biases 'prejudice'. We show here that prejudice is a special case of bias identifiable only by its negative consequences, and therefore impossible to eliminate purely algorithmically. Rather, prejudice requires deliberate action based on knowledge of a society and its outstanding ethical challenges.

If we are to demonstrate that AI incorporates the same bias as humans, we first have to be able to document human bias. We will use several methods to do this below, but the one we use most is the Implicit Association Test (IAT). First introduced by Greenwald et al. (1998), the IAT demonstrates enormous differences in response times when subjects are asked to pair two concepts they find similar, in contrast to two concepts they find different. The IAT follows a reaction time paradigm, which means subjects are encouraged to work as quickly as possible, and their response times are the quantified measure. For example, subjects are much quicker if they are told to label insects as unpleasant and flowers as pleasant than if they are asked to label these objects in reverse. The fact that a pairing is faster is taken to indicate that the task is more easy, and therefore that the two subjects are linked in their mind. The IAT is ordinarily used to pair *categories* such as 'male' and 'female' with *attributes* such as 'violent' or 'peaceful'. The IAT has been used to describe and account for a wide range of implicit prejudices and other phenomena, including stereotype threat (Kiefer and Sekaquaptewa, 2007; Stanley et al., 2011).

Our method for demonstrating both bias and prejudice in text is a variant of the implicit association test applied to a widely-used semantic representation of words in AI, termed *word embeddings*. These are derived by representing the textual context in which a word is found as a vector in a high-dimensional space. Roughly, for each word, its relationship to all other words around it is summed across its many occurrences in a text. We can then measure the distances (more precisely, cosine similarity scores) between these vectors. The thesis behind word embeddings is that words that are closer together in the vector space are semantically closer in some sense. Thus, for example, if we find that $\overrightarrow{\text{programmer}}$ is closer to $\overrightarrow{\text{man}}$ than to $\overrightarrow{\text{woman}}$, it suggests (but is far from conclusive of) a gender stereotype. We assume here that this measure is analogous to reaction time in the IAT, since the shorter time implies a semantic 'nearness' (McDonald and Lowe, 1998; Moss et al., 1995).

As with the IAT, we do not just compare two words. Many if not most words have multiple meanings, which makes pairwise measurements "noisy". To control for this, we use small baskets of terms to represent a concept. In the present paper we have never invented our own basket of words, but rather have in every case used the same words as were used in the psychological study we are replicating. We should note that distances / similarities of word embeddings notoriously lack any intuitive interpretation. But this poses no problem for us: our results and their import do not depend on attaching meaning to these distances. Our primary claim is that the associations revealed by relative nearness scores between categories match human biases and stereotypes strongly (i.e., low p-values and high effect sizes) and across many categories. Thus, the associations in the word vectors could not have arisen by chance, but instead reflect extant biases in human culture.